Psychometrika

A JOURNAL DEVOTED TO THE DEVELOPMENT OF PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE

THE PSYCHOMETRIC SOCIETY · ORGANIZED IN 1935

VOLUME 16
NUMBER 2
JUNE
1951

PSYCHOMETRIKA, the official journal of the Psychometric Society, is devoted to the development of psychology as a quantitative rational science. Issued four times a year, on March 15, June 15, September 15, and December 15

JUNE 1951 VOLUME 16, NUMBER 2

Printed for the Psychometric Society at 23 West Colorado Avenue, Colorado Springs, Colorado. Entered as second class matter, September 17, 1940, at the Post Office of Colorado Springs, Colorado, under the act of March 3, 1879. Editorial Office, Department of Psychology, The University of North Carolina, Chapel Hill, North Carolina.

Subscription Price: The regular subscription rate is $10.00 per volume. The subscriber receives each issue as it comes out, and a second complete set for binding at the end of the year. All annual subscriptions start with the March issue and cover the calendar year. All back issues are available. The price is $1.25 per issue or $5.00 per volume (one set only). Members of the Psychometric Society pay annual dues of $5.00, of which $4.50 is in payment of a subscription to *Psychometrika*. Student members of the Psychometric Society pay annual dues of $3.00, of which $2.70 is in payment for the journal.

Application for membership and student membership in the Psychometric Society, together with a check for dues for the calendar year in which application is made, should be sent to

RAYMOND A. KATZELL, Chairman of the Membership Committee
Psychological Services Center, Syracuse University, Syracuse 10, New York

Payments: All bills and orders are payable in advance. Checks covering membership dues should be made payable to the *Psychometric Society*. Checks covering regular subscription to *Psychometrika* and back issue orders should be made payable to the *Psychometric Corporation*. All checks, notices of change of address, and business communications should be addressed to

ROBERT L. THORNDIKE, Treasurer, Psychometric Society and Psychometric Corporation
Teachers College, Columbia University
New York 27, New York

Articles on the following subjects are published in *Psychometrika*:

(1) the development of quantitative rationale for the solution of psychological problems;
(2) general theoretical articles on quantitative methodology in the social and biological sciences;
(3) new mathematical and statistical techniques for the evaluation of psychological data;
(4) aids in the application of statistical techniques, such as nomographs, tables, work-sheet layouts, forms, and apparatus;
(5) critiques or reviews of significant studies involving the use of quantitative techniques.

The emphasis is to be placed on articles of type (1), in so far as articles of this type are available.

In the selection of the articles to be printed in *Psychometrika*, an effort is made to obtain objectivity of choice. All manuscripts are received by one person, who

(*Continued on the back inside cover page*)

Psychometrika

CONTENTS

(Continued)

VOLUME SIXTEEN JUNE 1951 NUMBER TWO

PSYCHOMETRIKA—VOL. 16, NO. 2
JUNE, 1951

A GENERAL SOLUTION FOR THE LATENT CLASS MODEL OF LATENT STRUCTURE ANALYSIS

BERT F. GREEN, JR.
EDUCATIONAL TESTING SERVICE
AND
PRINCETON UNIVERSITY

For the point distribution model of Lazarsfeld's latent structure analysis, the general matrix equation is stated which relates the manifest data in the form of joint occurrence matrices to the latent parameters. The relationship of the item responses and these joint occurrence matrices is also indicated in matrix form. A general solution for the latent parameters is then presented, which is based on the notion of factoring two joint occurrence matrices. The solution is valid under certain conditions which will usually be fulfilled. The solution assumes that estimates are available for the elements in the joint occurrence matrices with recurring subscripts, analogous to item communality or reliability. Some alternative methods of obtaining these estimates are discussed. Finally a fictitious 3-class, 8-item example is presented in detail.

Introduction

An important contribution to the theory of attitude measurement has recently been presented by Lazarsfeld* as part of a report of research in attitude measurement conducted for the Research Branch of the Information and Education Division of the Army Service Forces. Latent structure, as this innovation is called, is essentially a mathematical model for describing the interrelationships of items in an attitude questionnaire.

There are actually two latent structure models, one in which the underlying attitude variable and the item distributions are assumed to be continuous, and another in which the underlying attitude is assumed to have a point distribution. In the latter model, the individuals at a given point in the distribution are termed, collectively, a latent class. Both models deal with items having two response alternatives. If an item has more than two response categories, these may be combined in such a way that a response dichotomy is avail-

*Lazarsfeld, Paul. The logical and mathematical foundation of latent structure analysis. Stouffer, S., et al., *Studies in social psychology in World War II. Vol. IV. Measurement and Prediction*, Princeton: Princeton University Press, 1950.

able. One of the two alternative responses is arbitrarily designated as the positive response to the item. Latent structure analysis is concerned with the joint occurrence of such positive responses to items.

Lazarsfeld distinguishes between the manifest, or observable data, and the latent, or derived parameters of the model. The actual solution of the equations relating the manifest data to the latent parameters has been found only for certain special cases. Very little has been done for the continuous distribution model, and solutions have been reported only for special cases of the latent class model.* It is the purpose of the present article to present a general solution for the latent parameters for the latent class model, under a few conditions which will often be fulfilled.

The Latent Structure Equations

In the latent class model, each of the m points or latent classes is characterized by n_s^*, the proportion of people in latent class s, and v_{is}, the probability that a person in latent class s will respond positively to item i. If p_i is the proportion of people who respond positively to item i, we may write

$$p_i = \sum_{s=1}^{m} n_s v_{is} . \tag{1}$$

The fundamental hypothesis of latent structure analysis is that all item interrelationships may be completely explained by the mutual relationship of all items to the underlying attitude distribution. In the latent class model this means that the items are assumed to be independent for each class s, while the item intercorrelations are assumed to be due to the varying latent item probabilities in the different latent classes. Then, if p_{ij} is the proportion of people who respond positively to both items i and j, or the relative number of joint occurrences of positive responses to items i and j, we have

$$p_{ij} = \sum_{s=1}^{m} n_s v_{is} v_{js} . \tag{2}$$

In general, if $p_{ij\ldots k}$ is the proportion of people who respond positively to all of the set of items $i, j, \cdots k$,

*In connection with a RAND project at Columbia University, Mr. W. A. Gibson has devised a graphical method of obtaining approximate solutions if certain configurational criteria are met, while Dr. Lazarsfeld and Mr. J. Dudman have solved special cases by the use of asymmetric determinants.

$$p_{ij\ldots k} = \sum_{s=1}^{m} n_s \, v_{is} \, v_{js} \cdots v_{ks} \,. \tag{3}$$

Note finally that

$$\sum_{s=1}^{m} n_s = 1 \,. \tag{4}$$

Equations (1), (2), (3), and (4) express the manifest data as functions of the latent parameters n_s and v_{is}.

These equations may be written in matrix form. Following Lazarsfeld, with a few changes in notation, we define:

m = number of latent classes,
s = subscript designating latent class s; $s = 1, 2, \cdots m$,
r = number of items,
i = subscript designating item i; $i = 0, 1, 2, \cdots r$, where $i = 0$ has a special meaning defined in each case below,
σ = subscript denoting any subset of items, $i, j, \cdots k$,
N = $m \times m$ diagonal matrix, elements n_s,
L = $(r + 1) \times m$ matrix, elements v_{is}; $v_{os} = 1$, and
P_o = $(r + 1) \times (r + 1)$ symmetric matrix, elements p_{ij}; $p_{oi} = p_i$, $p_{oo} = 1$.

(Let p_{ii} remain undefined for the moment.)

$$P_o = \begin{Vmatrix} 1 & p_1 & p_2 & p_3 & \cdots & p_r \\ p_1 & p_{11} & p_{12} & p_{13} & \cdots & p_{1r} \\ p_2 & p_{21} & p_{22} & p_{23} & \cdots & p_{2r} \\ p_3 & p_{31} & p_{32} & p_{33} & \cdots & p_{3r} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ p_r & p_{1r} & p_{2r} & p_{3r} & \cdots & p_{rr} \end{Vmatrix}$$

D_k = $m \times m$ diagonal matrix, elements v_{ks}, k fixed; i.e., the $(k+1)$th row of L is made into a diagonal matrix.
P_k = $(r + 1) \times (r + 1)$ symmetric matrix, elements p_{ijk}, k fixed; $p_{ojk} = p_{jk}$, $p_{ook} = p_k$.
P_{kh} = $(r + 1) \times (r + 1)$ symmetric matrix, elements p_{ijkh}; k, h fixed; $p_{ojkh} = p_{jkh}$, $p_{ookh} = p_{kh}$.

Equations (1) to (4) are equivalent to

$$P_o = L\,N\,L'. \tag{5}$$

$$P_k = L\,N\,D_k\,L'. \tag{6}$$

$$P_{kh} = L\,N\,D_k\,D_h\,L'. \tag{7}$$

In general, we may write

$$P_\sigma = L\,N\,D_\sigma\,L' \quad \text{where } D_\sigma = \prod_{(g)} D_g\,,\ g \text{ ranging over the items in } \sigma\,, \text{ and } D_o = I\,. \tag{8}$$

Now define

$$A = L\,N^{\frac{1}{2}}, \tag{9}$$

so (8) becomes

$$P_\sigma = A\,D_\sigma\,A'. \tag{10}$$

Since $v_{os} = 1$, it follows from (9) that $a_{os} = \sqrt{n_s}$, while in general $a_{is} = v_{is}\sqrt{n_s}$. Thus it is sufficient to determine A, since from A, N and L are readily calculable.

Two further matrices will be needed. Define

$$D_{(1)} = \sum_{k=1}^{r} D_k\,. \tag{11}$$

The ith diagonal element in $D_{(1)}$ is the sum of all the elements but v_{os} in the sth column of L. Also define

$$P_{(1)} = \sum_{k=1}^{r} P_k = \sum_{k=1}^{r} A\,D_k\,A' = A\left(\sum_{k=1}^{r} D_k\right)A' = A\,D_{(1)}\,A'. \tag{12}$$

The matrix $P_{(1)}$ contains all the triple order joint occurrence proportions p_{ijk} in symmetric fashion. The use of such a matrix symmetric in all the items was suggested by Dr. Paul Horst.

The matrix equations imply manifest terms of the type p_{ii}, p_{iij}, p_{iii}, etc., terms in which the same item appears more than once in the subscript. We have called these terms the elements with recurring subscripts. These elements are actually not observable, but, for consistency, are considered to be defined by the general matrix equation (10). In the present solution it is assumed that estimates are available for these elements with recurring subscripts. Some methods of obtaining these estimates are discussed below.

Since the joint occurrence matrices form the basis for the solution for A, it is of theoretical and practical interest to state in

matrix notation the relationship of the basic item response data to these matrices. We define

$X = (r + 1) \times n$ matrix, elements $x_{i\alpha}$.

$x_{i\alpha}$ { $= 1$ if individual α responds positively to item i. / $= 0$ otherwise.

$x_{o\alpha} = 1$ for all α.

$U_k = n \times n$ diagonal matrix, elements $x_{k\alpha}$, k fixed; i.e., a row of X is made into a diagonal matrix.

$$U_{(1)} = \sum_{k=1}^{m} U_k .$$

The αth diagonal element in $U_{(1)}$, denoted $u_{\alpha(1)}$, is the number of items to which individual α responds positively. Now let

$$n\hat{P}_o = X X' ; \tag{13}$$

$$n\hat{P}_k = X U_k X' . \tag{14}$$

In general,

$$n\hat{P}_\sigma = X U_\sigma X', \quad \text{where } U_\sigma = \prod_{(g)} U_g , \; g \text{ ranging over the items in } \sigma , \text{ and } U_o = I . \tag{15}$$

$\hat{P}_\sigma$ is equivalent to P_σ except for the elements with recurring subscripts. Also let

$$n\hat{P}_{(1)} = n \sum_{k=1}^{r} \hat{P}_k = \sum_{k=1}^{r} X U_k X' = X \left(\sum_{k=1}^{r} U_k \right) X' = X U_{(1)} X'. \tag{16}$$

Finally, we have

$$P_o = \hat{P}_o + E_o , \tag{17}$$

where E_o is an $(r + 1) \times (r + 1)$ diagonal correction matrix with diagonal elements

$$e_{ii(o)} = p_{ii} - p_i ;$$
$$e_{oo(o)} = 0 .$$

Also,

$$P_{(1)} = \hat{P}_{(1)} + E_{(1)} , \tag{18}$$

where

$E_{(1)}$ is a correction matrix of order $(r+1) \times (r+1)$, with elements $e_{ij(1)} = p_{iij} + p_{ijj} - 2p_{ij}$, $(i \neq j)$;

$$e_{ii(1)} = \left(\sum_{k=1}^{r} p_{iik} \right) - \hat{p}_{ii_{(1)}}, \quad \left(\hat{p}_{ii_{(1)}} \text{ is a diagonal element of } \hat{P}_{(1)} \right);$$

$$e_{oi(1)} = p_{ii} - p_i;$$

$$e_{oo(1)} = 0.$$

From equation (16) it is evident that the element in the ith row and jth column of $\hat{P}_{(1)}$ is the sum of the $u_{a(1)}$ for those people who respond positively to both items i and j. This can readily be obtained with an IBM sorter and tabulator.

The Solution for the Latent Parameters

The conditions under which the present solution is valid are that $m \leq r + 1$, that A be non-singular, and that all the diagonal elements of $D_{(1)}$ be different and non-zero. That is, A must have no fewer rows than columns, the columns of A must be linearly independent, and the sums of the latent parameters for the different latent classes must be different and non-zero. In addition, as we have said, it is necessary to know the elements with recurring subscripts.

Under these conditions it follows from equations (10) and (12) that P_o and $P_{(1)}$ are of rank m. We may now operate on these matrices in the following manner.

From (10)

$$P_o = A\,A'. \tag{19}$$

By any factor method, as Thurstone's or Hotelling's, we may obtain

$$P_o = B\,B'. \tag{20}$$

From (19) and (20) and since P_o is of rank m, we must have

$$B\,\Lambda_b = A, \text{ where } \Lambda_b \text{ is orthogonal}. \tag{21}$$

Again by any factor method, we obtain

$$P_{(1)} = C\,C'. \tag{22}$$

From (12)

$$P_{(1)} = A\,D_{(1)}\,A' = (A\,D_{(1)}^{\frac{1}{2}})\,(A\,D_{(1)}^{\frac{1}{2}})'. \tag{23}$$

From (22) and (23), and since $P_{(1)}$ is of rank m, we must have

$$C\,\Lambda_c = A\,D_{(1)}^{\frac{1}{2}} \quad \text{where } \Lambda_c \text{ is orthogonal}. \tag{24}$$

From (21) and (24),

$$B\,\Lambda_b\,D_{(1)}^{\frac{1}{2}} = C\,\Lambda_c; \tag{25}$$

$$B\,\Lambda_b\,D_{(1)}^{\frac{1}{2}}\,\Lambda_c' = C\,. \tag{26}$$

Let

$$T \equiv \Lambda_b\,D_{(1)}^{\frac{1}{2}}\,\Lambda_c'. \tag{27}$$

From (26) and (27), we may solve for T in

$$B\,T = C\,. \tag{28}$$

The least-squares solution for T, found by minimizing the trace of $(BT - C)'(BT - C)$, is

$$T = (B'\,B)^{-1}\,B'\,C\,. \tag{29}$$

This method of obtaining T was pointed out by Dr. Paul Horst. From (27)

$$T\,T' = \Lambda_b\,D_{(1)}\,\Lambda_b'. \tag{30}$$

Since $D_{(1)}$ is diagonal, and since its diagonal elements are all different and non-zero, it is unique except for order, and contains the characteristic roots of $T\,T'$. Λ_b is also unique except for order, which is unimportant in this model. Thus a complete principal component analysis of $T\,T'$ will yield Λ_b, which, when premultiplied by B, will yield A.

The method of solution is to factor P_o and $P_{(1)}$, obtaining B and C respectively. From B and C obtain T by equation (29). Perform a complete principal component analysis of $T\,T'$ to obtain Λ_b. Obtain the product $B\,\Lambda_b$ which is A, the matrix containing the latent structure parameters.

After the foregoing was written, a simplification of the solution was pointed out by Dr. T. W. Anderson. From equation (29),

$$T\,T' = (B'\,B)^{-1}\,B'CC'B\,(B'B)^{-1}. \tag{30a}$$

We may substitute in (30a) from equation (22) yielding

$$T\,T' = (B'\,B)^{-1}B'P_{(1)}B\,(B'B)^{-1}. \tag{30b}$$

This implies that $P_{(1)}$ need not actually be factored. This solution involves factoring P_o to obtain B, and obtaining TT' by the matrix multiplication indicated in (30b). Λ_b is obtained in the same manner in both methods of solution.

An advantage of the factor analysis approach to the problem is that the number of classes can be determined in the factoring pro-

cedure. In most other approaches to the problem, the number of classes to be used must be estimated before starting the computations.

Since P_o and $P_{(1)}$ are factored in obtaining the solution, the derived latent parameters may be expected to fit these matrices. However, the solution makes no use of the higher-order joint occurrence frequencies p_{ijkl}, etc. Thus, in order to be sure that the derived parameters fit the complete set of data, all joint occurrence matrices would have to be checked against their computed counterparts. In practice, a check of some joint occurrence frequencies picked at random would probably suffice to determine whether the latent class model could be used to summarize the data. For fallible data or estimated unknowns, the characteristic roots of TT' will not be exactly the same as the sums of columns of L. The fit of the derived latent parameters is determined in part by this discrepancy as well as by the discrepancies of actual and computed joint occurrence matrices.

It should be noted that any higher order P_σ could be used in place of $P_{(1)}$ in the solution, if its corresponding D_σ satisfied the conditions imposed on $D_{(1)}$. It was felt that while many D_k might not satisfy these conditions, $D_{(1)}$ might ordinarily be expected to. Furthermore, since $P_{(1)}$ is symmetric in the items, it yields a unique solution.

This solution is not ideal. The development has been almost entirely in terms of algebra, with no statistical estimation procedures implied, nor any sampling theory. It would be desirable to find a solution which would be assured of fitting all the joint occurrence matrices in some "best" statistical sense, and which did not require estimating the elements with recurring subscripts.

Estimation of Elements with Recurring Subscripts

The elements with recurring subscripts in latent structure analysis are akin to the communalities in factor analysis; the estimation problems are similar in some respects. However, the necessity of estimating p_{iij} and p_{iii} poses additional complications. From the definition of $P_{(1)}$, it can be seen that there are many terms involved in it which must be estimated. While there are $r(r+1)^2$ terms represented in $P_{(1)}$, since each element represents the sum of r terms, some of these terms are duplicates. There are actually $(r^3+5r)/6$ different terms observable, and r^2+r different terms to be estimated. For small numbers of items, there are about as many terms to be

estimated as to be observed; in fact for $r < 7$, there are actually more terms to be estimated. It is clear then that we must either have good estimation procedures or restrict ourselves to fairly large number of items.

From equations (1) to (4) it is clear that limiting values for p_{ii} are

$$p_i \geq p_{ii} \geq p_i^2. \tag{31}$$

It can be shown that if the latent class probabilities for items i and j, v_{is} and v_{js}, are proportional, then the ith and jth columns of P_o are proportional. If we take the first row of P_o as its first factor, using Thurstone's diagonal method, then the same type of proportionality applies to the first residual matrix, $\|p_{ij} - p_ip_j\|$. If $(v_{is} - v_{it})$ is proportional to $(v_{js} - v_{jt})$ for all s and t, then the columns in the first residual matrix for items i and j are proportional. It is suggested that P_o and the first residual matrix be inspected for proportional columns, and diagonal entries inserted to maintain the proportionalities. Otherwise it might be reasonable to take the highest $(p_{ij} - p_ip_j)$ as an estimate of $(p_{ii} - p_i^2)$.

After P_o has been factored, using an estimate of p_{ii}, a new estimate of p_{ii} may be computed from the factor matrix. If this is quite different from the original estimate, and if the number of items is small, it may be best to refactor P_o using the new estimate. However, for larger numbers of items, the discrepancies probably will not markedly affect the factor matrix.

Having the estimate of p_{ii} from the factor matrix of P_o, we may proceed to the estimation of p_{iij} and p_{iii}. From equations (1) to (4) it is clear that

$$p_{ij} \geq p_{iij}. \tag{32}$$

Since all principal minors of P_j must be non-negative, we have, taking the oth and ith rows and columns,

$$\begin{vmatrix} p_j & p_{ij} \\ p_{ij} & p_{iij} \end{vmatrix} \geq 0. \tag{33}$$

From (32) and (33) we have, as limits for p_{iij},

$$p_{ij} \geq p_{iij} \geq \frac{p^2_{ij}}{p_j}. \tag{34}$$

By thinking of $p_{ii|j}$ as a probability, we may write

$$p_{ii|j} = p_{ii|j}\, p_j\,, \tag{35}$$

where $p_{ii|j}$ is the conditional probability of ii given j. One plausible assumption is that

$$\frac{p_{ii|j}}{p_{ii}} = \frac{p_{i|j}}{p_i}. \tag{36}$$

From this we obtain for an estimate of p_{iij},

$$p_{iij} = \frac{p_{ii}\, p_{ij}}{p_i}. \tag{37}$$

It may be verified that

$$2(p_{iij}\, p_i - p_{ii}\, p_{ij}) = \sum_{s,t=1}^{r} n_s\, n_t\, v_{is}\, v_{it}(v_{js} - v_{jt})(v_{is} - v_{it}). \tag{38}$$

Then $p_{iij} > \dfrac{p_{ii}\, p_{ij}}{p_i}$ if $v_{is} \gtrless v_{it}$ according as $v_{js} \gtrless v_{jt}$, which would seem to be the case in which i and j are positively related. If i and j are negatively related, probably $p_{iij} < \dfrac{p_{ii}\, p_{ij}}{p_i}$. From this it is suggested, in intuitive fashion only, that the estimate of p_{iij} be taken as

$$p_{iij} = \frac{p_{ii}\, p_{ij}}{p_i}(1 + p_{ij} - p_i p_j). \tag{39}$$

Similarly, for p_{iii} we have

$$p_{ii} \geq p_{iii} \geq \frac{p_{ii}^2}{p_i}. \tag{40}$$

If

$$\frac{p_{ii|i}}{p_{ii}} = \frac{p_{i|i}}{p_i}, \quad \text{then} \quad p_{iii} = \frac{p_{ii}^2}{p_i}; \tag{41}$$

or

$$p_{iii} = \frac{p_{ii}^2}{p_i}(1 + p_{ii} - p_i^2). \tag{42}$$

In estimating p_{iij} and p_{iii} by equations (37) or (39), and (41) or (42), if the estimates fall outside the range of possible values, as given in (34) and (40), the nearer limit should be taken as the estimate.

When the elements with recurring subscripts have been estimated, the method of solution can be applied to the data to obtain the latent parameters. From these obtained parameters new estimates of the unknown elements may be computed, using the equations implied in (10), i.e.,

$$p_{ii} = \sum_{s=1}^{m} n_s v_{is}^2; \tag{43}$$

$$p_{iij} = \sum_{s=1}^{m} n_s v_{is}^2 v_{js}; \tag{44}$$

$$p_{iii} = \sum_{s=1}^{m} n_s v_{is}^3. \tag{45}$$

If these values are somewhat different from the original estimates, the entire solution may be recomputed using the new estimates. From these calculations, a second set of derived latent parameters are available, from which a third set of estimates of the unknowns may be calculated. Using these, a third set of latent parameters may be derived. This iterative procedure should be continued until the estimates of the unknowns found from the last set of derived parameters are about the same as the estimates computed from the previous set of parameters.

It may be remarked that in theory it is possible to actually compute the unknown elements if the ranks of P_o and P_j are known. One would proceed by using the properties of "links" and "basic" determinants as discussed by Lazarsfeld.* Equations involving asymmetric determinants with one unknown element could be set up. However, it seems likely that such a procedure would be quite cumbersome.

These methods for estimating the elements with recurring subscripts are offered as suggestions rather than as final statements. With more experience in using the model, and with the gradual accumulation of empirical results, it is hoped that investigators in the field will devise better practical estimation procedures.

*Lazarsfeld, Paul, *op. cit.*

Illustrative Example

In order to illustrate the method of solution, a fictitious 3-class, 8-item example has been prepared. The hypothetical latent structure is presented in N_o and L_o', Table 1. (For convenience N_o is written

TABLE 1

	N_o	L_o 0	1	2	3	4	5	6	7	8	$\sum_{i=1}^{8}$
I	.2	1.0	.1	.6	.5	.2	.0	.0	.1	.3	1.8
II	.3	1.0	.5	.7	.0	.7	.7	.0	.2	.9	3.7
III	.5	1.0	.9	.8	.7	.2	.8	.5	.4	.5	4.8

as a column vector.) In Tables 2 and 3 are presented the joint occurrence matrices P_o and $P_{(1)}$, respectively. These matrices contain

TABLE 2
P_o

	0	1	2	3	4	5	6	7	8
0	1.000	.620	.730	.450	.350	.610	.250	.280	.580
1	.620	.482	.477	.325	.199	.465	.225	.212	.366
2	.730	.477	.539	.340	.251	.467	.200	.214	.425
3	.450	.325	.340	.295	.090	.280	.175	.150	.205
4	.350	.199	.251	.090	.175	.227	.050	.086	.251
5	.610	.465	.467	.280	.227	.467	.200	.202	.389
6	.250	.225	.200	.175	.050	.200	.125	.100	.125
7	.280	.212	.214	.150	.086	.202	.100	.094	.160
8	.580	.366	.425	.205	.251	.389	.125	.160	.386

the true values of the p_{ii}, p_{iij}, and p_{iii}. However, since these values are unknown in any practical situation, estimated values have been substituted for these true values throughout the calculations.

The first factor of P_o was obtained by Thurstone's diagonal method, pivoting on the first row of P_o. In the residual matrix, whose elements are $(p_{ij} - p_i p_j)$, the highest off-diagonal entry in a column

TABLE 3
$P_{(1)}$

	0	1	2	3	4	5	6	7	8
0	3.870	2.751	2.913	1.860	1.329	2.697	1.200	1.218	2.307
1	2.751	2.225	2.138	1.530	.828	2.117	1.080	.979	1.590
2	2.913	2.138	2.210	1.452	.971	2.080	.960	.945	1.724
3	1.860	1.530	1.452	1.266	.372	1.344	.840	.690	.894
4	1.329	.828	.971	.372	.654	.928	.240	.355	.961
5	2.697	2.117	2.080	1.344	.928	2.080	.960	.923	1.659
6	1.200	1.080	.960	.840	.240	.960	.600	.480	.600
7	1.218	.979	.945	.690	.355	.923	.480	.432	.691
8	2.307	1.590	1.724	.894	.961	1.659	.600	.691	1.532

was taken as the estimate of $(p_{ii} - p_i^2)$, except for the first two columns where the property of proportionality was used. After two more factors were extracted by the centroid method the residuals were vanishingly small; (the frequency distribution of residuals had mean $= 0$, standard deviation $= .003$). From the factor loadings, new estimates of the diagonal elements were obtained, which in turn were used to recompute the factor loadings. The factor matrix B_1' is presented in Table 4. Note that we did not have to start with knowledge of the number of latent classes; the factor analysis procedure indicated clearly that three classes were sufficient.

TABLE 4
B_1'

	0	1	2	3	4	5	6	7	8
I	1.000	.620	.730	.450	.350	.610	.250	.280	.580
II	0	.268	.067	.255	−.169	.164	.252	.120	−.100
III	0	.157	.041	−.137	.173	.249	.022	.048	.208

The p_{ii} computed from B_1 were used to estimate the p_{iij} and p_{iii} by means of equations (34) and (37). Using these estimates in $P_{(1)}$, we obtained the factor matrix C_1' by the diagonal and centroid methods, Table 5. Following the outlined procedure, T_1, Δ_{b1}, β_1, N_1,

TABLE 5
C_1'

	0	1	2	3	4	5	6	7	8
I	1.967	1.397	1.481	.941	.679	1.368	.611	.620	1.175
II	0	.466	.122	.530	−.358	.266	.504	.199	−.237
III	0	.265	.052	−.261	.261	.381	.032	.064	.316

and L_1' were obtained. Tables 6, 7, and 8 present these. (β_1 is a vector containing the characteristic roots of TT_1'.)

TABLE 6
T_1

1.9648	−.0016	−.0185
.4390	1.9306	−.0049
.3862	−.2280	1.6207

TABLE 7
Λ_{b1}

.733	.430	.526
.656	−.651	−.382
.178	.625	−.760

β_1

4.806	3.620	2.183

TABLE 8
L_1'

N_1	0	1	2	3	4	5	6	7	8	$\sum_{i=1}^{8}$
.537	1.000	.898	.799	.645	.240	.816	.480	.400	.542	4.820
.185	1.000	.442	.688	−.135	.858	.723	−.100	.167	1.035	3.678
.277	1.000	.198	.622	.462	.222	.131	.036	.124	.352	2.147

From N_1 and L_1, new estimates of the unknowns were available. Since these estimates were considerably different from the original estimates, both sets of factors were recomputed, and N_2 and L_2 obtained. At this point the estimates of the unknown elements indicated that the factors of P_o could not be improved, while the factor load-

ings from $P_{(1)}$ might be improved. Thus $B_2' = B_3'$, Table 9, while P_1 was refactored to obtain C_3', Table 10. From these factor matrices, T_3, Λ_{b3}, β_3, N_3 and L_3' were computed and are presented in Tables 11, 12, and 13.

TABLE 9

$B_2' = B_3'$

	0	1	2	3	4	5	6	7	8
I	1.000	.620	.730	.450	.350	.610	.250	.280	.580
II	0	.269	.068	.258	−.164	.170	.250	.117	−.098
III	0	.157	.042	−.143	.168	.254	.018	.044	.206

TABLE 10

C_3'

	0	1	2	3	4	5	6	7	8
I	1.967	1.398	1.481	.942	.677	1.372	.610	.619	1.174
II	0	.467	.121	.553	−.379	.248	.477	.215	−.265
III	0	.238	.063	−.217	.257	.389	.028	.070	.313

TABLE 11

T_3

1.9650	.0006	−.0002
.4450	1.9485	.0021
.3820	−.3476	1.5256

TABLE 12

Λ_{b3}

.706	.526	.474
.704	−.595	−.388
.079	.608	−.790

β_3

4.817	3.737	1.896

TABLE 13

N_3	L'_3									$\sum_{i=1}^{8}$
	0	1	2	3	4	5	6	7	8	
.498	1.000	.905	.803	.691	.205	.807	.501	.402	.506	4.820
.277	1.000	.498	.702	−.008	.730	.711	−.011	.198	.930	3.750
.225	1.000	.137	.603	.477	.205	.049	.015	.112	.316	1.914

A comparison of L_0' and L_3' indicates a close agreement. The necessity of the iterations in this example is probably due to the small number of items. The more items, the smaller the relative error in P_o and $P_{(1)}$ due to the estimation of unknowns, so the closer B_1 and C_1 would be to their true values.

It should be noted that in some cases the roots of TT' may be very similar, causing computational difficulties. It will usually be possible to spread out the values of these roots by reversing the positive and negative response designations for some items. Since the roots correspond to sums of columns of L, an inspection of the values in the first trial L will indicate which items should be changed. (If the scoring of an item is reversed, $P_{(1)}$, must be recalculated, but for P_o, if the first row has been taken as the first factor, the elements in the corresponding row and column of the residual matrix, elements $(p_{ij} - p_i p_j)$, are merely reversed in sign. Hence the centroid factor loadings for that item are merely reversed in sign, while the first factor loading is changed from p_i to $1 - p_i$.)

Summary

The general formulation of latent structure analysis is presented, following Lazarsfeld, for the case in which the underlying attitude variable is assumed to have a point distribution. For this case, called the latent class model, the general matrix equation, (10), relating the manifest data to the latent parameters is stated. Also stated is the general matrix equation, (15), which relates the basic item response data to the joint occurrence matrices. Under the restrictions that $m \leq r + 1$ (where m is the number of latent classes and r is the number of items), that A, (9), is non-singular, that $D_{(1)}$, (11), has diagonal entries all different and non-zero, and that the elements with recurring subscripts are known, a solution of equation (10) is presented: equations (20), (22), (29), (30), and (21). Some methods of obtaining estimates for the joint occurrences with recurring subscripts, such as p_{ii}, p_{iij} and p_{iii}, are discussed. Finally a fictitious 8-item example is presented in detail.

It is hoped that further work at a theoretical level will disclose better solutions than the one presented here, and that a solution will be found for the potentially more powerful continuous distribution model.

Manuscript received 7/26/50.
Revised manuscript received 10/5/50.

PSYCHOMETRIKA—VOL. 16, NO. 2
JUNE, 1951

TIME-LIMIT TESTS: ESTIMATING THEIR RELIABILITY AND DEGREE OF SPEEDING

LEE J. CRONBACH AND W. G. WARRINGTON
UNIVERSITY OF ILLINOIS

Non-spurious methods are needed for estimating the coefficient of equivalence for speeded tests from single-trial data. Spuriousness in a split-half estimate depends on three conditions; the split-half method may be used if any of these is demonstrated to be absent. A lower-bounds formula, r_c, is developed. An empirical trial of this coefficient and other bounds proposed by Gulliksen demonstrates that, for moderately speeded tests, the coefficient of equivalence can be determined approximately from single-trial data. It is proposed that the degree to which tests are speeded be investigated explicitly, and an index τ is advanced to define this concept.

Introduction

Most group tests of mental ability and achievement are administered with time limits. In some cases, the time limits are of no importance, as nearly every subject completes all he can do correctly. In other tests, the limits are short enough to make rate of work an important factor in the score. Careful distinction between these types of time-limit test, has been lacking, and all time-limit tests have generally been referred to as speed tests.

It is a maxim that the "reliability" of speed tests should not be determined by single-trial procedures such as the split-half and Kuder-Richardson formulas. This prohibition is presented in an unqualified fashion in many texts, an example being this statement (6, p. 69):

> In a speeded test, it is impossible to determine a coefficient of equivalence save by giving an immediate parallel test. The split-half and Kuder-Richardson methods must not be used. When this principle is disregarded, as it often is, and split-half methods are applied to speed tests, the resulting coefficients are grossly inflated.

Comparable statements are made by Thorndike (16, pp. 86, 112), Adkins (1, p. 152), and Guilford (9, pp. 486, 496), who do, however, indicate incidentally that some time-limit tests are so little speeded that the internal-consistency methods may be used.

Although such prohibitions have been voiced repeatedly, they are as commonly violated. Many test authors, even in recent manuals, report reliability estimates based on procedures known to give spuriously high results. Some introduce additional misinterpretation, thus: "The reliabilities . . . were computed by means of the Kuder-Richardson formula, which may underestimate the true reliability but should never over-estimate it." (12). So well-informed a pair of investigators as the Thurstones, in a major factorial study (15), base all their reliability estimates on K-R Formula 20 (we shall call this coefficient α in the rest of this paper), although many of the tests are speeded.

The prevalence of spurious coefficients and their interference with test evaluation is suggested by the repeated criticism by Buros' reviewers on this point (3, pp. 347, 410, 459, 518, 532, 630-631, 636). While the tone of the review varies, each one echoes, in effect, the words of E. K. Taylor (3, p. 631): "Since the use of the Kuder-Richardson formulae on speed tests is entirely inappropriate, the reliability of the instrument remains an unknown quantity."

Essentially we have a situation where the test user is given reliability coefficients which—according to accepted principles—are no good at all, and which if anything do harm by making the tests look more dependable than they are. Why does such a situation arise? Partly we can blame authors and publishers for lack of scruple. There is temptation to report spurious coefficients, since such data make tests more salable. But we must also recognize that almost prohibitive labor is involved in getting a non-spurious estimate of reliability by usual methods. The Thurstones, for example, would have had to double the amount of testing in their factorial study, to get better coefficients.

We cannot ignore the probability that disregard of the commandment against spurious procedures stems (in part) from the sweeping nature of the pronouncement. The injunction is frustrating to authors and publishers because it blocks an easy technique. Then consumers are educated to demand coefficients, and no reasonably economical method for supplying them is sanctioned. Worst of all, there is considerable evidence that the pronouncement is unreasonable and unjustified in many instances. Where a test is only slightly speeded, and so most people finish all they can do, split-half or α coefficients are *not* misleading. In many instances where both single-trial and two-trial coefficients have been determined, the single-trial coefficients are not noticeably higher.

Whatever the causes, we lack interpretable reliability coefficients

for most time-limit tests. To relieve the situation, we can (a) establish limits within which it is safe to rely on conventional single-trial analyses for speeded tests or (b) develop usable single-trial procedures which do not give spurious results. Gulliksen (10) has taken steps in this direction, and offers several suggestions. We shall, in this paper, demonstrate mathematically the nature and causes of spuriousness in single-trial estimates and derive a lower-bounds formula to correct for spuriousness. Then we shall compare results of several procedures, including Gulliksen's, applied to a set of tests which vary as to speeding. We shall develop formally the concept of degree of speeding. Finally, we arrive at working recommendations for dealing with speeded tests.

A Rationale

The reliability coefficient is supposed to indicate the stability of a test score over a period of time, or the equivalence of two forms of a test (7). The stability of a performance from day to day or month to month cannot possibly be estimated from single-trial data. Ordinarily we are concerned with obtaining a coefficient of equivalence, a measure of the consistency of the person's standing on two measurements at the same time. If only one form of a test is available, the split-half method or Kuder-Richardson formula α is used to indicate the consistency of measurement.

But to determine the equivalence of two measures, we need two or more sets of independent data. The parallel-form method, for example, requires that the two tests be administered independently. When we split a test in order to get two sets of scores, we are willing to assume that the many items within an unspeeded test are independent of one another. That is, especial difficulty on one of the items neither increases nor decreases the person's probable standing on the remainder. But in a timed test, the person who gets stuck on one item may never reach the remainder of the items. It is this interdependence of items that introduces spuriousness. We turn now to defining the conditions for spuriousness mathematically.

Suppose a test containing n items ($i = 1, 2, \cdots, n$) is taken by N persons ($p = 1, 2, \cdots, N$). Then, on any trial, each person completes f_p items in the time allowed. If person p attempts item i, he earns a score x_{ip} on the item. His test score is x_p.

$$x_p = \sum_{i=1}^{f_p} x_{ip}. \tag{1}$$

In order to account for the possibility that some persons finish well before time is called, we envision "invisible" items at the end of the test, all of which are so difficult that everyone's score on them is inevitably zero. This device permits f_p to be greater than n, and permits f_p to be greater for a person who finishes the test early than for one who finishes late.

When many independent trials on the same items are given, we estimate increasingly accurately the person's true number of items finished $f_{p\infty}$ and the true score on any item (when attempted) $x_{ip\infty}$.

$$f_{p\infty} = \bar{f}_p \text{ and } x_{ip\infty} = \bar{x}_{ip} .$$

On some particular trial,

$$f_p = f_{p\infty} + \nu_p ,$$

and if item i is finished by person p (i.e., $i \leq f_p$),

$$x_{ip} = x_{ip\infty} + \varepsilon_{ip} .$$

ν and ε are error terms and may be positive or negative.

$$x_p = \sum_{i=1}^{f_{p\infty}+\nu_p} (x_{ip\infty} + \varepsilon_{ip}) = \sum_{1}^{f_{p\infty}} x_{ip\infty} + \sum_{f_{p\infty}}^{f_{p\infty}+\nu_p} x_{ip\infty} + \sum_{1}^{f_{p\infty}} \varepsilon_{ip} + \sum_{f_{p\infty}}^{f_{p\infty}+\nu_p} \varepsilon_{ip} . \quad (2)$$

Now the desired coefficient of equivalence is the correlation between scores on two sets of items, administered independently. We shall indicate parameters of the second test by primes (thus, x'_p, f'_p, etc). To make the tests comparable, we assume $x_{ip\infty} = x'_{ip\infty}$, $M_x = M_{x'}$, $\sigma_x = \sigma_{x'}$, $f_{p\infty} = f'_{p\infty}$.

$$x'_p = \sum_{1}^{f_{p\infty}} x_{ip\infty} + \sum_{f_{p\infty}}^{f_{p\infty}+\nu'_p} x_{ip\infty} + \sum_{1}^{f_{p\infty}} \varepsilon'_{ip} + \sum_{f_{p\infty}}^{f_{p\infty}+\nu'_p} \varepsilon'_{ip} , \quad (3)$$

$$r_{xx'} = \frac{\frac{1}{N}\sum_{1}^{N} x_p x'_p - M_x M_{x'}}{\sigma_x \sigma_{x'}} ,$$

$$r_{xx'} = \frac{\frac{1}{N}\sum x_p x'_p - M_x^2}{\sigma_x^2} . \quad (4)$$

To examine the effect of violation of the assumption of independence upon the correlation between tests, we expand the term $\sum xx'$. To simplify notation, we drop the subscript p hereafter. Multiplying (2) by (3) and summing,

$$\sum_{1}^{N} xx'$$

$$= \sum_{p}\Bigg[\underset{\text{(a)}}{(\sum_{i=1}^{f_\infty} x_{i\infty})^2} + \underset{\text{(b)}}{\sum_{1}^{f_\infty} x_{i\infty} \sum_{f_\infty}^{f_\infty+v'} x_{i\infty}} + \underset{\text{(c)}}{\sum_{1}^{f_\infty} x_{i\infty} \sum_{1}^{f_\infty} \varepsilon'_i} + \underset{\text{(d)}}{\sum_{1}^{f_\infty} x_{i\infty} \sum_{f_\infty}^{f_\infty+v'} \varepsilon_i}$$

$$+ \underset{\text{(e)}}{\sum_{f_\infty}^{f_\infty+v} x_{i\infty} \sum_{1}^{f_\infty} x_{i\infty}} + \underset{\text{(f)}}{\sum_{f_\infty}^{f_\infty+v} x_{i\infty} \sum_{f_\infty}^{f_\infty+v'} x_{i\infty}} + \underset{\text{(g)}}{\sum_{f_\infty}^{f_\infty+v} x_{i\infty} \sum_{1}^{f_\infty} \varepsilon'_i} + \underset{\text{(h)}}{\sum_{f_\infty}^{f_\infty+v} x_{i\infty} \sum_{f_\infty}^{f_\infty+v'} \varepsilon'_i}$$

$$+ \underset{\text{(i)}}{\sum_{1}^{f_\infty} \varepsilon_i \sum_{1}^{f_\infty} x_{i\infty}} + \underset{\text{(j)}}{\sum_{1}^{f_\infty} \varepsilon_i \sum_{f_\infty}^{f_\infty+v'} x_{i\infty}} + \underset{\text{(k)}}{\sum_{1}^{f_\infty} \varepsilon_i \sum_{1}^{f_\infty} \varepsilon'_i} + \underset{\text{(l)}}{\sum_{1}^{f_\infty} \varepsilon_i \sum_{f_\infty}^{f_\infty+v'} \varepsilon'_i}$$

$$+ \underset{\text{(m)}}{\sum_{f_\infty}^{f_\infty+v} \varepsilon_i \sum_{1}^{f_\infty} x_{i\infty}} + \underset{\text{(n)}}{\sum_{f_\infty}^{f_\infty+v} \varepsilon_i \sum_{f_\infty}^{f_\infty+v'} x_{i\infty}} + \underset{\text{(o)}}{\sum_{f_\infty}^{f_\infty+v} \varepsilon_i \sum_{1}^{f_\infty} \varepsilon'_i} + \underset{\text{(p)}}{\sum_{f_\infty}^{f_\infty+v} \varepsilon_i \sum_{f_\infty}^{f_\infty+v'} \varepsilon'_i} \Bigg]. \qquad (5)$$

We assume that $\sum_{p=1}^{N} v_p = 0$ and $\sum_{p=1}^{N} \varepsilon_{ip} = 0$ for any trial. (Throughout this analysis, we are dealing with population statistics, and N is accordingly large). We also assume that v_p and ε_{ip} are independent of each other and of $f_{p\infty}$ and $x_{ip\infty}$.

Whether the two tests are experimentally independent or not, terms (c), (g), (h), (i), (j), (l), (n), and (o) reduce to zero as a consequence of the definition of error as independent of true score, and of the relation $\sum_{p} \varepsilon_{ip} = 0$. We also assume that $r_{\varepsilon_i \varepsilon'_j} = 0$, hence $(k) = 0$. This is the assumption involved in the usual split-half method applied to unspeeded tests, namely, that departures from true score on one item are independent of departures on other items, when both items are reached by the subject.

$$\sum xx' = \sum_{p} (a + b + d + e + f + m + p). \qquad (6)$$

If x *and* x′ *are independent,* $r_{vv'} = 0$ and $(f) = 0$. Then

$$\sum xx'_{\text{independent}} = \sum (a + b + d + e + m + p).$$

And, since $\sum(b) = \sum(e)$, and $\sum(d) = \sum(m)$,

$$\sum xx'_{\text{independent}} = \sum(a + 2b + 2d + p). \qquad (7)$$

If x *and* x′ *are not independent,* being obtained from items administered within a single time limit, $r_{vv'} \neq 0$. Then

$$\sum xx'_{\text{spurious}} = \sum(a + b + d + e + f + m + p).$$

Collecting symmetrical terms,

$$\sum xx'_{\text{spurious}} = \sum (a + 2b + 2d + f + p). \tag{8}$$

Subtracting (7) from (8),

$$D = \sum xx'_{\text{spurious}} - \sum xx'_{\text{independent}} = \sum_p (f) = \sum_p \sum_{f_\infty}^{f_\infty + \nu} x_{i\infty} \sum_{f_\infty}^{f_\infty + \nu'} x_{i\infty} . \tag{9}$$

D indicates the spuriousness of a single-trial estimate. The difference between r_s the spurious single-trial estimate of equivalence, and r_i the comparable-form coefficient, is a monotonic function of D. The right-hand member of (9) is a covariance which depends on the magnitude of ν, the correlation between ν and ν', and the size of the $x_{ip\infty}$ for items such that $f_\infty - |\nu| < i < f_\infty + |\nu|$. D approaches zero if

(a) ν approaches zero, i.e., there is little variation from trial to trial in number of items finished by a person;

or

(b) $r_{\nu\nu'}$ approaches zero, ie., fluctuations from true number of items finished on one test are independent of fluctuations on the other test;

or

(c) the $x_{ip\infty}$ for items the person is working on as time runs out approach zero.

As all persons come close to completing the test, variation in ν approaches zero, and in this case D approaches zero.

Conditions required for single-trial estimates to be acceptable without correction. If *any one* of the following conditions is satisfied, spuriousness is negligible and the single-trial methods (split-half or α) may be used.

(1) The variation in number of items finished is small.

(2) When time is called, subjects have completed all the items which they have an appreciable chance of getting right. (Or, if a correction formula is used in scoring the test, subjects have completed all items where their probability of success is greater than chance.)

(3) Fluctuations in number finished from trial to trial are small, compared to variation in number finished from person to person.

Single-trial data do not permit us to estimate fluctuations in number finished. But $\sigma_v < \sigma_f$, and we may sometimes demonstrate that condition (1) holds by showing that σ_f approaches zero. To demonstrate condition (2), we may compute the success of each person on the last one or two items attempted; if this approaches zero, condition (2) is satisfied. These methods provide a basis for demonstrating, under certain circumstances, that a single-trial estimate of reliability is a good estimate of the value to be obtained from independent forms. On the other hand, we cannot conclude that the estimate is necessarily spurious in cases where neither condition (1) nor (2) obviously obtains. The above approach leaves something to be desired, for it is expressed in the indefinite words "small," "appreciable," and the like.

Estimating Bounds to the True Coefficient

If we can obtain an upper bound for D, we can correct the spurious coefficient of equivalence (r_s) to get a lower bound for the true coefficient from independent trials (r_i).
From (4),

$$r_i = r_s - \frac{D}{N\sigma^2_x}, \tag{10}$$

$$D = \sum_p \sum_{f_\infty}^{f_\infty+v} x_{i\infty} \sum_{f_\infty}^{f_\infty+v'} x_{i\infty}. \tag{9}$$

We now assume that

$$\sum_{f_\infty}^{f_\infty+v} x_{i\infty} \leq x_{\infty f_\infty} \cdot v, \tag{11}$$

where $x_{\infty f_\infty}$ is the person's true score on the f_∞th item. This assumption becomes true when items are arranged in order of difficulty, and the true order of difficulty of items is the same for each person. Should the latter condition not be true, departures from (11) will nevertheless average out, over the group of subjects.

$$D \leq \sum_1^N x^2_{\infty f_\infty} vv'. \tag{12}$$

Now there are N_a cases for whom $f_\infty \leq n$, and for them, the term within the summation of (12) is not zero. For the remaining cases

$(p_{a+1} \cdots p_N)$ whose last item is one of the "invisible" items beyond the nth, $x_{\infty f_\infty}$ is zero and the term within the summation is zero. Therefore,

$$D \leq \sum_{1}^{N_a} x^2_{\infty f_\infty} \nu\nu' . \tag{13}$$

Now, assume $\nu\nu' = \nu^2$, i.e., $r_{\nu\nu'} = 1.00$. This overestimates the degree of spuriousness. Since ν is independent of $x_{\infty f_\infty}$,

$$D < \frac{1}{N_a} \sum_{1}^{N'_a} x^2_{\infty f_\infty} \sum_{1}^{N_a} \nu^2 . \tag{14}$$

If we let $\sigma_{\nu(a)}$ represent the standard deviation of ν for the first a cases,

$$D < \sigma^2_{\nu(a)} \sum_{1}^{N_a} x^2_{\infty f_\infty} , \tag{15}$$

$$r_i > r_s - \frac{\sigma^2_{\nu(a)} \sum_{1}^{N_a} x^2_{\infty f_\infty}}{N\sigma^2 x} . \tag{16}$$

In order to estimate this, we make use of the relation $\sigma_\nu < \sigma_f$,

$$r_i > r_s - \frac{\sigma^2_{f(a)} \sum_{1}^{N_a} x^2_{\infty f_\infty}}{\sigma^2_x N} . \tag{17}$$

Since N_a is defined in terms of true number finished, single-trial data permit us only to make approximations to the above correction. We may assume, however, that the number of cases whose *true* number finished is n or fewer is approximately the same as the number whose *obtained* number finished is n or fewer (where, as before, we define number finished in terms of the "invisible" items). Our data do not show clearly whether the person who reached the end of the test did so with time to spare, so that his last item is one of the invisible ones, or whether he barely finished. We propose to estimate the number of cases who finished the nth item, but did not finish the $(n+1)$th invisible item, by assuming that this number equals the number who completed the $(n-1)$th item but not the nth. (A more elaborate method, involving extrapolation from the several preceding items, seems unnecessary.) We therefore determine $\sigma^2_{f(a)}$ from the frequency distribution of number of items finished, where the frequency at n items is set equal to the frequency at $n-1$ items.

Estimating $\sum_{1}^{N_a} x^2_{\infty f_\infty}$ is slightly more difficult. Neither $f_{p\infty}$ nor

$x_{ip\infty}$ is known. We therefore make our estimate in terms of $x_{\lambda p}$, the person's mean obtained score on the last items finished. Because the mean true score, over many persons, equals the mean obtained score, and because the variance of obtained scores is greater than the variance of true scores,

$$\sum_{1}^{N_a} x^2{}_{\infty f_\infty} < \sum_{1}^{N_a} x^2{}_{\lambda p}. \tag{18}$$

$x_{\lambda p}$ might be estimated from only the last item, but this would add considerable error variance to the estimate. As more items are averaged to get $x_{\lambda p}$, the estimate becomes closer to the true score, making the members of (18) less unequal. But as still more items are added, so that the easier items nearer the start of the test are included, the estimated $x_{\lambda p}$ again becomes larger and the inequality more extreme. We therefore have estimated $x_{\lambda p}$ by averaging each person's score on the last two items reached. Because we wish to sum over N_a cases, the value of $x^2{}_{\lambda p}$ is obtained for each person who reaches the end of the test, and averaged. This value is then multiplied by the number of persons reaching only n—1 items, and entered in the total together with the values of $x^2{}_{\lambda p}$ for persons completing n—1 or fewer items.

$$r_i > r_c = r_s - \frac{\sigma^2{}_{f(a)} \sum_{1}^{N_a} x^2{}_{\lambda p}}{N \sigma^2{}_x}. \tag{19}$$

The desired coefficient of equivalence, r_i lies between r_s and r_c. Because a great many inequalities are introduced in deriving (19), the lower bound will often be far from r_i. There seems to be no purpose in obtaining a lower bound which is far below the upper bound, so that the coefficient is essentially undetermined. But in instances where D, as estimated, is small, the upper and lower bounds will be close together and r_i can be inferred satisfactorily from single-trial data. It is therefore recommended that equation (19) be used in evaluating single-trial data for time-limit tests. Our derivation depends upon a large number of cases, and large samples should be used in practical work with the formula. If the test is essentially unspeeded, r_s and r_c will be close together and a confident report regarding the coefficient of equivalence can be made. If the bounds are widely separated, no useful conclusion is possible.

The computational procedures leading to r_c are not involved. The steps are as follows:

1. Determine number of items finished (f_p) for each person.

2. Make a frequency distribution for the number of persons having each number of items finished, entering the value for n—1 items opposite n items also. From this, compute $\sigma^2_{f(a)}$.

3. For the next step, errors should be marked on the answer sheet with a colored pencil. For each person, average his score on the last two items and square the value obtained. (This square will be .00, .25, or 1.00.) If N_0 persons complete the test, and N_1 persons complete all but the last item, sum the $x^2\lambda_p$ for the N_0 persons and multiply by N_1/N_0. To this value, add the $x^2\lambda_p$ for all persons not finishing the test.

4. Enter these values in (19), together with σ_x^2 obtained in the usual fashion, to get r_c.

Gulliksen's Formulas and Modifications of Them

Gulliksen (10) has recently derived three lower-bounds formulas intended to serve as single-trial estimates of reliability for speeded tests. These formulas are:

$$r_S = r_s - \frac{\sigma_f^2}{\sigma_x^2}; \tag{20}$$

$$r_M = r_s - \frac{n - M_f}{\sigma_x^2}; \tag{21}$$

$$r_K = r_s - \frac{n(n - M_f) - (n - M_f)^2 - \sigma_f^2}{(n-1)\,\sigma_x^2}. \tag{22}$$

Gulliksen states that $r_K > r_M > r_S$, but since variance in unfinished may be small compared to mean unfinished, this is a relation which tends to be true rather than a mathematical necessity. In these formulas, f refers to number of items finished, with no reference to the hypothetical "invisible" items included within f in our rationale.

Gulliksen recognizes that a better lower bound might be obtained if n, the total number of items, were replaced by an estimate of the number of items actively differentiating between persons. Items which everyone reaches, and items which no one reaches, play no part in the variance of f. We have developed this suggestion into two additional formulas, each of which would be expected to give a better lower bound than $r_{K'}$. If k is the greatest number of items completed by any subject, and if m is the smallest number completed by any subject,

$$r_{K'} = r_s - \frac{k(k - M_f) - (k - M_f)^2 - \sigma_f^2}{(k-1)\,\sigma_x^2}, \tag{23}$$

and

$$r_{K''} = r_s - \frac{(k-m)(k-m-M_f) - (k-m-M_f)^2 - \sigma_f^2}{(k-m-1)\,\sigma_s^2}. \quad (24)$$

$$r_{K''} \geq r_{K'} \geq r_K .$$

From (21), we have the similarly derived formula,

$$r_{M'} = r_s - \frac{k - M_f}{\sigma_s^2}, \quad (25)$$

$$r_{M'} \geq r_M .$$

Since the precise value of k, and of m, depends on only a single case, formulas (23), (24), and (25) are subject to marked fluctuation from sample to sample. As in the case of r_c, $r_{K'}$ and $r_{K''}$ should be based on large samples.

Gulliksen's initial formulation assumes that error variance on the test (E_s^2, in our notation) may be broken into two independent portions, E_Y^2 based on items marked incorrectly and E_U^2 (our ν^2) based on items left unfinished. His statement,

$$E_s^2 = E_Y^2 + \nu^2 , \quad (26)$$

disregards the probable negative correlation between E_Y and ν. In correspondence, he points out that such a correlation probably has little effect on the reliability estimate. In any case the more correct statement,

$$E_s^2 < E_Y^2 + \nu^2 , \quad (27)$$

is not inconsistent with Gulliksen's formulas. The second point to be mentioned is that, in deriving r_M, Gulliksen employs the relation $\nu^2 < M_U$ (U being the number of items unfinished, $n-f$). In deriving this he implicitly assumes that the correlation between reaching items g and h within the same trial ($r_{g_f h_f}$) is negligibly greater, over all item-pairs, than the correlation when the items are independently administered ($r_{g_f h'_f}$). Gulliksen's derivation of r_K is based on a similar assumption. Correspondence with Gulliksen clarifies the point that he justifies such an assumption on the empirical grounds that any spuriousness of the type $r_{g_f h_f} > r_{g_f h'_f}$ has a negligible effect on the reliability coefficient. This seems probable at first glance, and our empirical data tend to support this view. The Gulliksen formulas would rest on a sounder base if the conditions required for this assumption to hold were examined.

Empirical Analysis

We have made an empirical analysis of a set of test data in order to answer the question: Just how inaccurate are single-trial estimates of reliability for speeded tests? Much of this work was done before the foregoing rationale was developed, but in any case empirical evidence is needed to demonstrate forcefully to test authors what risk they run of misleading readers when they publish such coefficients. The analysis also permits us to try the newly proposed lower-bounds formulas, to evaluate their usefulness.

The data treated were made available by Dr. Merle Tate. He administered four mental tests individually, noting the time required by each person to complete each item (14). The 36 high-school students used as subjects were directed to work for both speed and accuracy. A maximum time of three minutes was allowed for any item, but this limit was rarely reached. He had four tests, ranging from 60 to 64 items, dealing with arithmetic reasoning, number series, sentence completion (vocabulary) and spatial relations. His items were drawn from typical mental tests, but his experimental procedure may make the results somewhat different from results obtainable in ordinary group testing.

We obtain a test score for each student for any given time limit. By determining the cumulative time on successive items, we determine how many items he would have finished by 1200 seconds, for example; then we count how many of those he solved correctly.

The "true" coefficient of equivalence r_i was obtained by dividing the test into comparable halves, and adding the times for each student as if the two halves had been administered independently. Guttman's split-half formula (11) was used to obtain the reliability. The spurious split-half reliability r_s was obtained when the two half-tests were timed within a single time-limit. Results for some of the time-limits employed are presented in Table 1 and graphed in Figure 1. (The times used for Test IV differ from the other tests because our procedure was changed after treating that set of data.) It should be noted that our sample is small, so that our results are markedly influenced by sampling error.

These conclusions follow:

1. Under some circumstances, single-trial split-half estimates are much higher than the true coefficient. This is markedly seen in Tests III and IV, with short time limits. ***Split-half single-trial estimates cannot be usefully interpreted, unless they are also accompanied by evidence that the degrees of spuriousness is negligible.***

2. The spuriousness declines as the time is extended, and is negligible even when some students have not finished the test.

3. For the present tests, spuriousness is small when the time-limit allows an average of 30-40 seconds or more per item.

4. For Tests I and II, the single-trial estimate is not markedly spurious even when very short time-limits are imposed.

The Kuder-Richardson formulas have been used by some test authors in the belief that these formulas lead to underestimates and so will not give spuriously high coefficients. But the factors that cause a split-half coefficient to be spuriously high also operate in the Kuder-Richardson formulas, and in fact, α is the mean of all possible split-half coefficients for the given test (5). When the Kuder-Richardson formulas were applied in the usual manner to the Tate tests, results shown in Table 2 were obtained.

TABLE 1

True and Spurious Coefficients of Equivalence with Varying Time Limits

	Time (seconds)	200	400	800	1600	2400	3200	4800
Test I Arithmetic Reasoning	r_s	.795	.846	.893	.918	.915	.874	.866
	r_i	.790	.810	.889	.892	.904	.876	.857
Test II Number Series	r_s	.906	.895	.856	.875	.888	.888	.881
	r_i	.795	.824	.766	.834	.894	.884	.881
Test III Sentence Completion	r_s	.864	.913	.923	.879	.883	.877	.877
	r_i	.574	.699	.838	.861	.883	.877	.877
	Time (Seconds)	600	900	1200	1500	2400	3300	4800
Test IV Spatial	r_s	.860	.880	.890	.922	.938	.924	.911
	r_i	.539	.574	.756	.764	.890	.911	.911

TABLE 2
Coefficients Estimated by the Kuder-Richardson Method and by Various Lower-Bound Formulas

	Time (seconds)	200	400	800	1600	2400	3200	4800
	r_i	.790	.810	.889	.892	.904	.876	.857
	r_c*	—	—	—	—	—	—	—
Test I	KR20 (α)	.587	.732	.818	.856	.855	.827	.813
Arithmetic	KR21	.142	.450	.708	.800	.824	.791	.766
Reasoning	r_M	−7.784	−2.507	−.122	.609	.806	.823	.862
	r_K	−.361	.076	.507	.736	.839	.835	.862
	$r_{K'}$	.177	.456	.671	.736	.839	.835	.862
	$r_{K''}$	.583	.629	.771	.800	.873	.852	.866
	r_i	.795	.824	.766	.834	.894	.884	.881
	r_c	.077	.277	.342	.558	.823	.883	.881
Test II	KR20 (α)	.765	.789	.806	.797	.782	.775	.770
Number	KR21	.135	.190	.603	.708	.721	.721	.718
Series	r_M	−3.437	−1.819	.024	.621	.823	.886	.881
	r_K	−.028	.002	.446	.691	.833	.875	.881
	$r_{K'}$	.480	.413	.520	.691	.833	.875	.881
	$r_{K''}$	.586	.511	.632	.742	.853	.882	.881
	r_i	.574	.699	.838	.861	.883	.877	
	r_c	.266	.314	.504	.830	.883	.877	
Test III	KR20 (α)	.770	.826	.855	.833	.822	.809	
Sentence	KR21	.387	.635	.781	.811	.810	.811	
Completion	r_M	−2.774	−.233	.576	.827	.879	.877	
	r_K	.137	.500	.718	.837	.879	.877	
	$r_{K'}$	.527	.714	.725	.837	.879	.877	
	$r_{K''}$	.622	.792	.800	.858	.881	.877	

	Time (seconds)	600	900	1200	1800	2400	3600	4800
	r_i	.539	.574	.756	.880	.890	.911	.911
	r_c	.267	.337	.411	.705	.819	.902	.911
	KR20 (α)	.812	.868	.878	.921	.925	.910	.901
Test IV	KR21	.492	.679	.733	.863	.891	.882	.871
Spatial	r_M	−1.094	−.021	.337	.753	.856	.905	.911
	r_K	.261	.509	.606	.821	.878	.907	.911
	$r_{K'}$	.550	.690	.730	.821	.878	.907	.911
	$r_{K''}$	.669	.754	.782	.869	.901	.913	.911

*These values were not computed because items in Test I are not arranged in approximate order of difficulty.

The spuriousness is sufficiently great for Tests III and IV to demonstrate that α is no more defensible as a single-trial estimate than is the split-half method. The KR-21 coefficient is lower than the

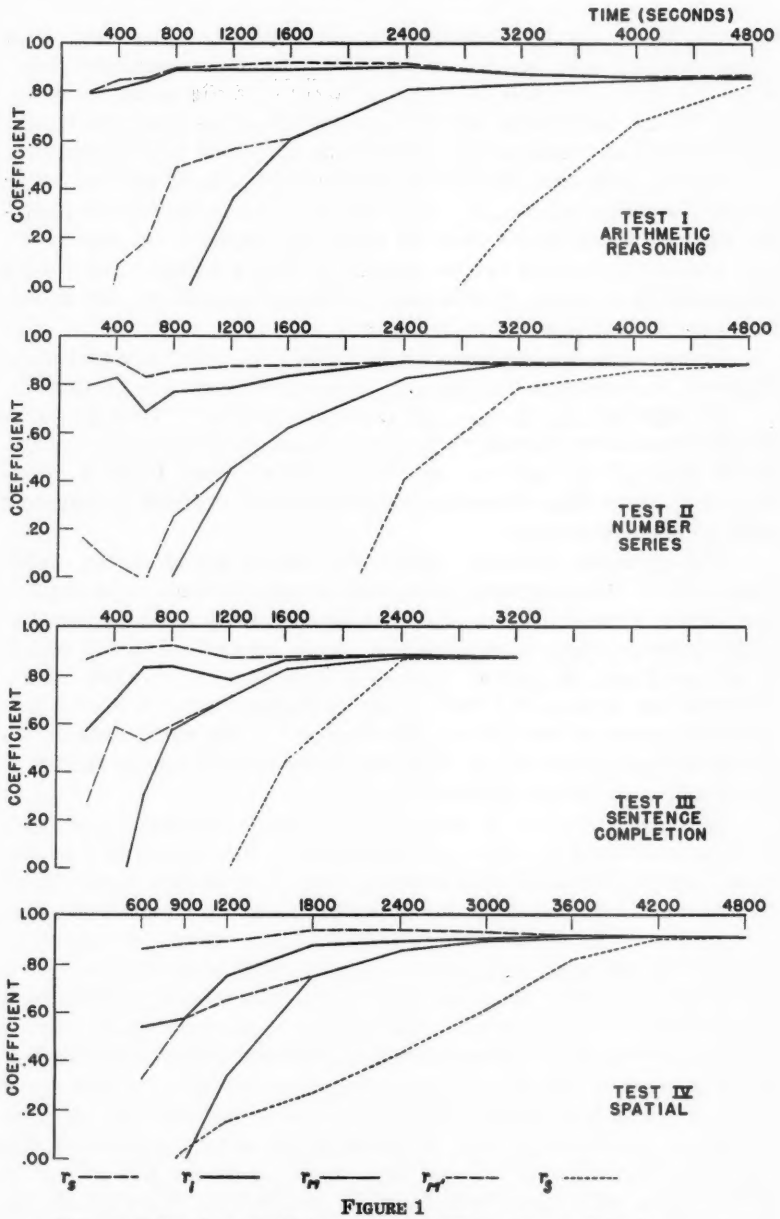


FIGURE 1

True and Spurious Coefficients, with Gulliksen Lower Bounds, as a Function of Time Allowed.

true coefficient for all tests and for all degrees of speeding, with minor exceptions in Test IV. KR-21 is not to be recommended, however. We have been able to construct tests of quite usual tpyes for which KR-21 coefficients are spuriously high. The good results obtained with Tate's data are therefore to be dismissed as a coincidence.

Table 2 also presents results from application of several lower bounds formulas: r_c, r_M, r_K, $r_{K'}$, and $r_{K''}$. In practical test analysis, samples much larger than 36 cases are required for use of r_c, $r_{K'}$, and $r_{K''}$. r_c could not be applied to Test I because items were not arranged in order of difficulty. Values of r_M and $r_{M'}$ are shown in Figure 1, and values of r_c in Figure 2.

Values of r_S are not included in Table 2, but they are plotted in Figure 1. r_S is always less than r_M, unless few items are finished.

As expected, r_S, r_M, $r_{M'}$, and r_c remain below r_i. They do therefore serve as lower bounds. r_K is a lower bound throughout the range, but in Test IV, $r_{K'}$ and $r_{K''}$ exceed r_i. This is very likely a consequence of using these formulas involving highly unstable parameters with so small a sample.

The spurious split-half coefficient always serves as an upper bound for r_i (disregarding the trivial instances where r_i is slightly the larger, presumably because of sampling error). Wherever one of the lower bounds comes close to r_s, r_i is determined to fall within a narrow range of values. r_S comes close to r_s only when nearly everyone has finished the test. r_M comes fairly close to r_s over a considerable range of time limits, and so does r_c. For short time limits, the lower bounds are so far from the upper bounds that r_i cannot be estimated with useful accuracy.

If r_M (or $r_{M'}$ or r_c) is close to r_s, r_i can be inferred. r_M is easier to determine than r_c, and will generally be the preferred formula, if we accept Gulliksen's assumption. However, r_M and r_c are based on different formulations of the problem, and for some tests they will be expected to give substantially different results. r_c takes into account the difficulty of the items being finished as time ends. r_c will therefore ordinarily be higher than r_M in tests where subjects have reached items which are quite difficult for them, and where there is considerable range in number finished. Whatever lower-bounds formula is used, if the lower-bound does not fall close to the upper bound, r_i cannot be estimated by a single-trial method with any useful degree of accuracy. It does not help, for example, to know that the lower bound is near .30 and the upper bound near .90. In test II, 200 seconds, with these bounds, r_i is .80; in Test IV, with similar bounds, r_i is only .54.

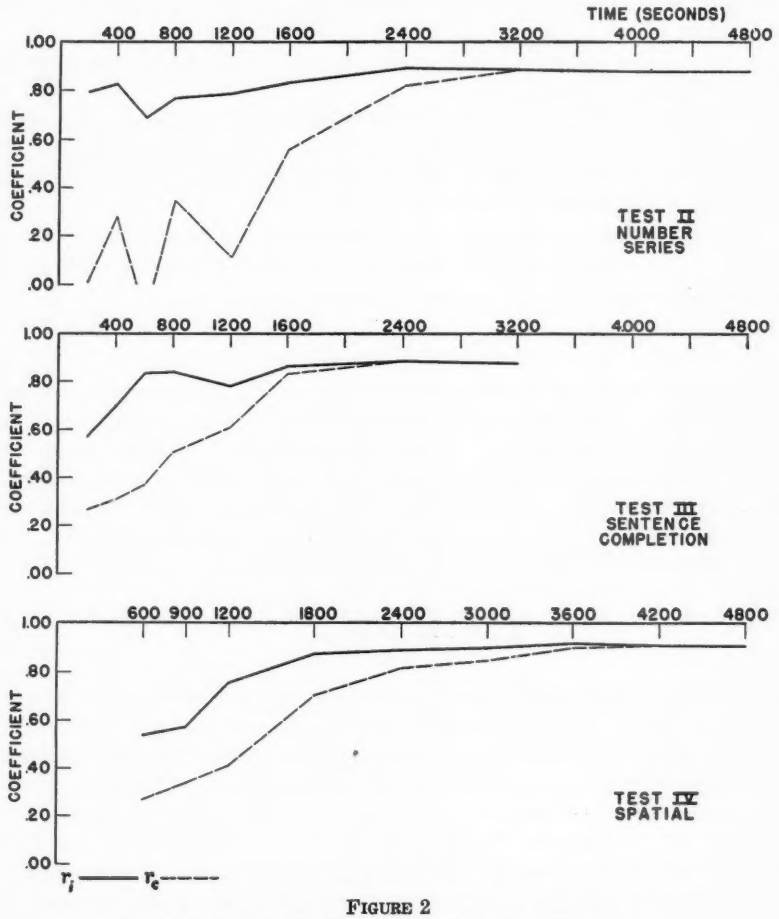


FIGURE 2
True Split-Half Coefficient with the Lower Bound r_c.

We evaluate the formulas as follows:

r_S: An extreme lower bound, easily computed. May be used as a rapid check to demonstrate that speeding is negligible. In case r_S falls much below r_s, another lower-bound formula should be used.

r_M, $r_{M'}$: Gives empirical results similar to r_c, and is much more readily computed. Involves an inadequately tested assumption, but

the probable soundness of the assumption justifies tentative use of r_M and $r_{M'}$.

r_c: Gives satisfactory empirical results, and can be used with r_s to locate r_i approximately.

r_K, $r_{K'}$, and $r_{K''}$: Values from these formulas, together with r_s, delimit r_i more narrowly than r_M or r_c. Adequate empirical trial impossible on present small sample. Formulas involve an assumption not yet adequately explored.

The test author is advised to compute r_s together with r_M, $r_{M'}$, or r_c. If the lower bound is close to r_s, r_i can be inferred with adequate accuracy.

A Note on Degree of Speeding

Test theory will be clarified if we can define and measure *degree of speeding*. Then the false dichotomy between speeded and unspeeded tests can be discarded. *A test is completely unspeeded when no subject's standing would be altered if he were given additional time.* Speeding is introduced if the time limit alters, not his score, but his true standard score in the group. A hint in an early study by Tinker and Paterson (13) led us to the index we call τ.

$$\tau = 1 - \frac{r_{A_t B_p} \cdot r_{A_p B_t}}{r_{A_t B_t} \cdot r_{A_p B_p}}. \tag{28}$$

A and B are equivalent forms of the test, and the subscripts t and p indicate scores under time-limit and power conditions respectively. In our study τ is estimated by correlating independently administered half-tests, the power condition being taken as the performance when all students had attempted all items. This index shows what proportion of the reliable variance in the score obtained with a given time limit reflects the same factor as the test does when given under unspeeded conditions. If τ is .90, ninety per cent of the true-score variance represents a speed factor, and only ten per cent represents whatever altitude factor is involved in responding to the test items.

Applied to Tate's tests, τ gives results shown in Table 3. In addition to the tabled data, results were also obtained for about an equal number of intermediate points. These conclusions follow:

1. The score variance due to speeding may be negligible even though many students have not finished. An index similar to τ is required for rigorous thinking about degree of speeding.

2. Insofar as these data are representative, speeding is very

TABLE 3
Degree of Speeding at Various Time Limits

Test	Statistic	Time (seconds) 200	400	800	1600	2400	3200	4800
I Arithmetic Reasoning	M_f	9	15	25	43	55	60	64
	M_x	6	10	15	23	28	31	33
	s_f	2.9	4.8	7.6	10.9	9.9	6.5	1.6
	s_x	2.5	3.8	6.2	8.3	9.1	8.5	8.1
	Tau	.64	.47	.34	.40	.32	.25	.02
II Number Series	M_f	13	20	31	48	57	59	60
	M_x	12	17	23	33	37	38	38
	s_f	3.5	4.4	7.3	7.7	4.8	2.3	0
	s_x	3.3	3.8	5.9	7.0	7.0	6.9	6.9
	Tau	.79	.69	.61	.29	.28	.00	0
III Sentence Completion	M_f	12	22	38	56	60	60	—
	M_x	10	18	28	37	38	38	—
	s_f	3.4	5.5	7.9	5.8	1.0	0	—
	s_x	3.6	5.8	8.0	8.4	8.3	8.3	—
	Tau	.49	.45	.25	.04	.00	0	—
	Time:	600	900	1200	1500	2400	3300	4800
IV Spatial	M_f	19	25	32	37	50	58	60
	M_x	14	19	23	27	35	38	39
	s_f	4.6	5.7	6.2	7.2	7.7	4.4	0
	s_x	4.6	6.2	7.2	8.7	10.9	10.5	9.7
	Tau	.42	.29	.17	.18	.12	.03	0

f = number finished
x = number right

slight with time limits averaging one minute per item for arithmetic reasoning and number series; or with time limits averaging thirty seconds per item for sentence completion and spatial items.

3. When τ is less than .05, the difference between r_s and r_i is negligible. If these empirical results are dependable, the single-trial coefficients are dependable for tests having τ .05 or less.

Gulliksen draws attention to the ratio $\frac{M_U}{s_x^2}$. This is also an index of speeding, which lacks the rational basis of τ but has the advantage of ease of computation. For the Tate tests, τ and the Gulliksen ratio give similar results, but in the long run the ratio is not especial-

ly useful because it makes no allowance for the difficulty of the items left unmarked.

General Recommendations

Our studies lead to these suggestions:

1. Split-half and Kuder-Richardson methods should not be used save to get an *upper bound* r_s for the coefficient of equivalence of a speeded test.

2. Demonstrating that either of two conditions (see text) holds for the test in question is a sufficient condition for assuming that the true coefficient is close to this upper bound.

3. Computing the lower bound r_c is useful. r_i is known to fall between r_s and r_c. If these are close together, r_i is usefully delimited. Instead of r_c, other lower bounds developed by Gulliksen's method may be advantageous, but they cannot be finally accepted because of the assumptions involved in the present derivations.

When a single-trial method is used, we can obtain only a coefficient of equivalence, showing how two samples of behavior taken while the person is operating under the same conditions would compare. Rates of work are probably unstable, and no single-trial coefficient can reveal how much of the "reliable" variance is due to temporary sets.

The analysis of Tate's data points to a recommendation for testing and test research in general. Both our analysis and Tate's show that variance on a speeded test contains separate portions attributable to speed and altitude. This conception was also developed earlier by Baxter (2) and Davidson and Carroll (8). This implies that degree of speeding may be an important characteristic of a test to be reported in test manuals (cf. 4). τ can be estimated fairly easily, provided two comparable forms can be administered. Knowing the degree of speeding would help significantly in interpreting a test.

In factorial studies and similar research, batteries should be designed in which speed factors would emerge along with altitude factors, as in the Davidson-Carroll study. The common practice of using substantially speeded tests in factorial batteries has probably caused us to ignore differences in factorial composition between, for example, speeded and unspeeded verbal tests. Factorial batteries designed to investigate relations between tests having various degrees of speeding would determine to what extent the speed variance represents the same factor in tests having different content, and to what extent the

speed variance represents ability rather than response set. Definition and measurement of the speed factor or factors would then lead naturally to a consideration of the role of speeding in test validity for various predictions.

The test author has responsibility for making available dependable information on the accuracy of his test. When a test is given with a time limit, there is no excuse for reporting a split-half coefficient alone. But the author may be able to report data, based on only a single administration of his test, which will satisfy the demand of the test consumer. If he can demonstrate that his test is only slightly speeded, or if he can establish the lower bound for the reliability by one of the formulas we have discussed (r_g, the most easily computed; r_c; or one of the Gulliksen bounds), he can report these single-trial estimates and need not use a two-trial method. But if the lower-bounds formulas give such low coefficients as to cast doubt on the usefulness of the test, two experimentally independent trials should be used to obtain a more exact coefficient of equivalence.

REFERENCES

1. Adkins, D. C. Construction and analysis of achievement tests. Washington, D. C.: U. S. Government Printing Office, 1947.
2. Baxter, B. An experimental analysis of the contribution of speed and level in an intelligence test. *J. educ. Psychol.*, 1941, 32, 285-296.
3. Buros, O. K. (Ed.) The third mental measurements yearbook. New Brunswick: Rutgers University Press, 1949.
4. Conrad, H. Information which should be provided by test publishers and testing agencies on the validity and use of their tests. Proceedings, 1949 Invitational Conference on Testing Problems, pp. 63-68. Princeton: Educational Testing Service, 1950.
5. Cronbach, L. J. Coefficient "alpha" and the internal structure of tests. To be published.
6. Cronbach, L. J. Essentials of psychological testing. New York: Harper and Brothers, 1949.
7. Cronbach, L. J. Test "reliability": its meaning and determination. *Psychometrika*, 1947, 12, 1-16.
8. Davidson, W. M., and Carroll, J. B. Speed and level components in time-limit scores, a factor analysis. *Educ. psychol. Meas.*, 1945, 5, 411-427.
9. Guilford, J. P. Fundamental statistics in psychology and education. New York: McGraw-Hill Book Co., 1950.
10. Gulliksen, H. The reliability of speeded tests. *Psychometrika*, 1950, 15, 259-269.
11. Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, 10, 255-282.
12. Iowa Silent Reading Test, Manual. Yonkers: World Book Co., 1943.

13. Paterson, D. G., and Tinker, M. A. Time-limit *vs.* work-limit methods. *Amer. J. Psychol.*, 1930, 42, 101-104.
14. Tate, M. W. Individual differences in speed of response in mental test materials of varying degrees of difficulty. *Educ. psychol. Meas.*, 1948, 8, 353-374.
15. Thurstone, T. G., and Thurstone, L. L. Mechanical aptitude III: Description of group tests. Psychometric Laboratory Reports, No. 55, 1949.
16. Thorndike, R. L. Personnel selection. New York: John Wiley and Sons, Inc., 1949.

Manuscript received 8/10/50
Revised manuscript received 1/12/51

PSYCHOMETRIKA—VOL. 16, NO. 2
JUNE, 1951

OPTIMAL TEST LENGTH FOR MAXIMUM BATTERY VALIDITY

PAUL HORST
EDUCATIONAL TESTING SERVICE

Having given a fixed amount of total testing time it is important to know how long each test in the battery should be so that the correlation of the battery with the criterion will be a maximum. The precise solution for the test lengths will depend on a particular set of conditions which may be specified. The writer has previously presented solutions for two sets of conditions. This article presents the solution for a third set of conditions. These are: (1) The total number of items or testing time is fixed. (2) The score is the total number of items correctly answered. (3) The test lengths are determined in such a way that the correlation of total score with the criterion is a maximum. The solutions for the two previous sets of conditions, together with the current set, are summarized. A set of experimental data is submitted to each solution and the three sets of results are compared.

Suppose a battery of tests has been administered to an appropriate sample of persons on whom criterion measures are also available. We specify that a person's score on each test be simply the total number of items answered correctly. We assume now that the following data are available:

1. The number of items in each test.
2. The variance of each test.
3. The reliability of each test.
4. The validity of each test.
5. The intercorrelations of the tests.

Now if the lengths of the tests are altered, all of the statistics enumerated above will also be altered. In general we may assume the lengths of the tests to be somewhat arbitrary and we may wish to establish a rational basis for altering these lengths. Several rational bases may be suggested as follows:

1. The lengths of the tests should be such that the raw score multiple regression weights for predicting the criterion

measures are all equal. This means that the total number of items correct in all the tests would give the best least-squares prediction of the criterion.

2. Having specified a fixed number of total items for all tests, the lengths of the tests should be such that the multiple correlation of the tests with the criterion is a maximum. Here we make no restrictions as to the values of the regression weights and therefore we could get a higher multiple correlation than if the weights were all required to be equal.

The solution for determining the test lengths on the basis of unit regression weights has been presented elsewhere (1). The problem of determining the test lengths on the basis of maximum multiple correlation for fixed testing time has also been solved (2).

A third basis for altering test length has been proposed by Professor Harold Gulliksen, as follows. Having specified that (1) the total number of items desired is a fixed value, and (2) the score shall be the total number of items correct, let us alter the lengths of the original tests in such a way that the correlation of total score with the criterion shall be a maximum. At first it appears that Professor Gulliksen's conditions are the same as those given in the first rationale, namely, that the raw score regression weights should all be equal. However, it can be shown that the set of conditions proposed by Professor Gulliksen lead to a different solution. This we shall proceed to show. First, however, we shall indicate the solutions for the test lengths which satisfy the other two sets of conditions. For the sake of uniformity of treatment, these solutions will be presented in somewhat different form than in the references cited. For all three sets of conditions we specify that in altering test lengths we do not change the nature of the tests and hence preserve the following conditions:

1. The average item variance in the test remains unchanged.
2. The average inter-item covariance for the items within the test remains unchanged.
3. The average inter-item covariance between the items in one test and those in another test remains unchanged.
4. The average item-criterion covariance remains unchanged.

We let

a_i = the original length of test i,
b_i = the altered length of test i,
σ_i = the standard deviation of test i,
r_{ii} = the reliability of test i,
r_{ij} = the correlation of test i with test j, and
r_{ic} = the correlation of test i with the criterion.

To present the solutions for each of the three sets of conditions, we begin with the following four sets of simultaneous equations:

$$\left.\begin{aligned} r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1n}\beta_n &= r_{1c} \\ r_{12}\beta_1 + r_{22}\beta_2 + \cdots + r_{2n}\beta_n &= r_{2c} \\ \cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots & \\ r_{1n}\beta_1 + r_{2n}\beta_n + \cdots + r_{nn}\beta_n &= r_{nc} \end{aligned}\right\} \tag{1}$$

$$\left.\begin{aligned} r_{11}t_1 + r_{12}t_2 + \cdots + r_{1n}t_n &= \sigma_1(1 - r_{11}) \\ r_{12}t_1 + r_{22}t_2 + \cdots + r_{2n}t_n &= \sigma_2(1 - r_{22}) \\ \cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots & \\ r_{1n}t_1 + r_{2n}t_2 + \cdots + r_{nn}t_n &= \sigma_n(1 - r_{nn}) \end{aligned}\right\} \tag{2}$$

$$\left.\begin{aligned} r_{11}V_1 + r_{12}V_2 + \cdots + r_{1n}V_n &= \sqrt{a_1(1 - r_{11})} \\ r_{12}V_1 + r_{22}V_2 + \cdots + r_{2n}V_n &= \sqrt{a_2(1 - r_{22})} \\ \cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots & \\ r_{1n}V_1 + r_{2n}V_n + \cdots + r_{nn}V_n &= \sqrt{a_n(1 - r_{nn})} \end{aligned}\right\} \tag{3}$$

$$\left.\begin{aligned} r_{11}S_1 + r_{12}S_2 + \cdots + r_{1n}S_n &= a_1/\sigma_1 \\ r_{12}S_2 + r_{22}S_2 + \cdots + r_{2n}S_n &= a_2/\sigma_2 \\ \cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots & \\ r_{1n}S_1 + r_{2n}S_2 + \cdots + r_{nn}S_n &= a_n/\sigma_n \end{aligned}\right\} \tag{4}$$

The unknowns in equations (1) to (4) are the β's, t's, V's, and S's respectively. These unknowns have no special meaning other than in calculations. All of these sets of equations are similar in that the coefficients of the unknowns are the correlation coefficients among the test variables. The reliabilities of the tests are the coefficients in the diagonal positions. The right hand sides of the equations are all experimentally determined values but are different from one set of equations to the next. Since for each set of equations we have n equations in n unknowns, we can solve for the sets of unknowns by any one of a number of methods. Let us assume, then, that all of the β's, t's, V's, and S's have been solved for. We shall now show for each of the three sets of conditions how the altered test lengths, b_i, may be solved for in terms of these values and the appropriate values on the right hand sides of the equations. The proofs of the

solutions for the first two sets of conditions have been given in the references cited (1 and 2). The proof for the solution for the third set of conditions will follow in this article. The solutions are as follows.

Condition I. The tests shall be altered in length so that when the score on each test is total number of items correct, then the raw score regression weights shall all be equal. The new test lengths b are then*

$$\left.\begin{aligned} b_1 &= \frac{a_1}{\sigma_1}(C\beta_1 - t_1), \\ b_2 &= \frac{a_2}{\sigma_2}(C\beta_2 - t_2), \\ &\cdots\cdots \\ b_n &= \frac{a_n}{\sigma_n}(C\beta_n - t_n), \end{aligned}\right\} \tag{5}$$

where

$$C = \frac{\Sigma b_i + \Sigma \dfrac{a_i t_i}{\sigma_i}}{\Sigma \dfrac{a_i \beta_i}{\sigma_i}}. \tag{5a}$$

The multiple correlation in this case is given by

$$R^2 = \Sigma \beta_i r_{ic} - \frac{(\Sigma t_i r_{ic})\left(\Sigma \dfrac{a_i \beta_i}{\sigma_i}\right)}{2b_i + \Sigma \dfrac{a_i t_i}{\sigma_i}}, \tag{6}$$

where the multiple correlation we would get if all tests were perfectly reliable is given by the first term on the right of (6) namely, $\Sigma \beta_i r_{ic}$.

Condition II. The lengths of the tests shall be so altered that if the total length of the battery is fixed, the multiple correlation of the tests with the criterion shall be a maximum irrespective of the regression weights. Then the lengths of the tests will be

* These b's are not to be confused with raw score regression weights.

$$\left.\begin{aligned} b_1 &= \sqrt{a_1(1-r_{11})}\,(L\beta_1 - V_1), \\ b_2 &= \sqrt{a_2(1-r_{22})}\,(L\beta_2 - V_2), \\ &\cdots\cdots\cdots\cdots \\ b_n &= \sqrt{a_n(1-r_{nn})}\,(L\beta_n - V_n), \end{aligned}\right\} \tag{7}$$

where

$$L = \frac{\Sigma b_i = \Sigma V_i \sqrt{a_i(1-r_{ii})}}{\Sigma \beta_i \sqrt{a_i(1-r_{ii})}}. \tag{7a}$$

The multiple correlation for this case is given by

$$R^2 = \Sigma \beta_i r_{ic} - \frac{(\Sigma r_{ic} V_i)^2}{\Sigma b_i + \Sigma V_i \sqrt{a_i(1-r_{ii})}}. \tag{8}$$

In this case the a's and b's may be used to indicate testing time instead of number of items if we prefer.

Condition III. The lengths of the tests shall be so altered that if the total length of the battery is fixed and the score is the total number of items correct, the correlation of total score with the criterion shall be a maximum. This correlation must be equal to or less than for Case II since the weights are all equal. Then the lengths of the tests will be

$$\left.\begin{aligned} b_1 &= \frac{a_1}{\sigma_1}\left[\, G\beta_1 - \tfrac{1}{2}t_1 + KS_1 \,\right], \\ b_2 &= \frac{a_2}{\sigma_2}\left[\, G\beta_2 - \tfrac{1}{2}t_2 + KS_2 \,\right] \\ b_n &= \frac{a_n}{\sigma_n}\left[\, G\beta_n - \tfrac{1}{2}t_n + KS_n \,\right], \end{aligned}\right\} \tag{9}$$

where

$$G = \frac{\left(2\Sigma b_i + \Sigma \frac{t_i a_i}{\sigma_i}\right)^2 - \left[\Sigma (1-r_{ii})\sigma_i t_i\right]\left(\Sigma \frac{S_i a_i}{\sigma_i}\right)}{2\left[\left(2\Sigma b_i + \Sigma \frac{t_i a_i}{\sigma_i}\right)\left(\Sigma r_{ic} S_i\right) - \left(\Sigma r_{ic} t_i\right)\left(\Sigma \frac{S_i a_i}{\sigma_i}\right)\right]} \tag{9a}$$

and

$$K=\frac{\left[\Sigma(1-r_{ii})\sigma_i t_i\right](\Sigma r_{ic}S_i)-(\Sigma r_{ic}t_i)\left(2\Sigma b_i+\Sigma\frac{t_i a_i}{\sigma_i}\right)}{2\left[\left(2\Sigma b_i+\Sigma\frac{t_i a_i}{\sigma_i}\right)\left(\Sigma r_{ic}S_i\right)-\left(\Sigma r_{ic}t_i\right)\left(\Sigma\frac{S_i a_i}{\sigma_i}\right)\right]}. \quad (10)$$

In this case we do not get a multiple correlation in the strictest sense of the word. The correlation of total score with the criterion is given, however, by a rather complicated equation as follows. We let

$$\left.\begin{aligned}
&A_{11}=\Sigma r_{ic}\beta_i\,,\ A_{12}=\Sigma r_{ic}t_i\,,\ A_{13}=\Sigma r_{ic}V_i\,,\ A_{14}=\Sigma r_{ic}S_i\,,\\
&A_{22}=\Sigma t_i\sigma_i(1-r_{ii})\,,\ A_{23}=\Sigma t_i\sqrt{a_i(1-r_{ii})}\,,\ A_{24}=\Sigma t_i\frac{a_i}{\sigma_i},\\
&A_{33}=\Sigma V_i\sqrt{a_i(1-r_{ii})}\,,\ A_{34}=\Sigma V_i\frac{a_i}{\sigma_i},\\
&A_{44}=\Sigma S_i\frac{a_i}{\sigma_i}.
\end{aligned}\right\} \quad (11)$$

Then

$$R^2=\frac{\left[\begin{aligned}&A_{11}A_{22}A_{33}+2A_{12}A_{13}(A_{23}+\Sigma b_i)\\&-\left[A_{11}(A_{22}+2\Sigma b_i)^2+A_{22}A_{13}{}^2+A_{33}A_{13}{}^2\right]\end{aligned}\right]}{A_{22}A_{33}-(A_{23}+2\Sigma b_i)^2}. \quad (12)$$

It will be noted that equations (9) which give the lengths of the tests for the third set of conditions are not as simple as those for the first two sets of conditions. The b's for the first two sets require the solutions for only two sets of unknowns, while those for the third set require solutions for three sets of unknowns. Furthermore, two constants must be determined for conditions III, whereas only one is required for I and II. Also the solution for the constants G and K are considerably more complicated than for C and L. For all three sets of conditions, the constants are functions of Σb_i or the over-all specified length of the battery.

However, the same set of coefficients are used in solving for all four sets of unknowns, and only the constants on the right are different. If the Doolittle method for solving normal equations or some variant of it is used, then the forward solutions will all be identical

except for the constant columns and only the back solutions are distinctly different. But the back solutions are accomplished much more rapidly than the forward solutions. To illustrate the three different methods, we take data from tests administered for experimental purposes to 210 first-year students at Ohio State University Law School in the Fall of 1948. The tests are entitled

1. Verbal Analogies
2. Best Arguments
3. Practical Judgment

The data are as follows:

	Intercorrelations, r_{ij}			Validity, r_{ic}	Reliability, r_{ii}
	1	2	3		
1		.34	.01	.30	.78
2			−.05	.26	.52
3				.15	.12
σ_i	4.9	4.5	3.1		
a_i	3	10	10		

$$R^2 = .144\,.$$

Using these data in equations (1) through (4) we get

$\beta_1 = .122$	$t_1 = 2.603$	$V_1 = -3.301$	$S_1 = -3.885$
$\beta_2 = .562$	$t_2 = 8.392$	$V_2 = 9.132$	$S_2 = 9.811$
$\beta_3 = 1.472$	$t_3 = 26.403$	$V_3 = 28.761$	$S_3 = 31.254$

We shall also need the values given by equations (11). These are

$A_{11} = .404$	$A_{12} = 5.364$	$A_{13} = 5.701$	$A_{14} = 6.074$
	$A_{22} = 87.396$	$A_{23} = 94.687$	$A_{24} = 102.326$
		$A_{33} = 102.745$	$A_{34} = 111.157$
			$A_{44} = 120.361$

Assuming now we do not wish to change the total test length we have

$$\Sigma a = \Sigma b = 23\,.^*$$

For condition I we have

*This value is given in 5-minute units to facilitate computation. The total time is actually 115 minutes.

$$C = \frac{23 + 102.32}{6.08} = 20.612. \qquad \text{Eqs. (5a) \& (11)}$$

$$\left.\begin{aligned} b_1 &= .612(\ .122K + 2.60) = 3.13 \\ b_2 &= 2.222(\ .562K - 8.39) = 7.10 \\ b_3 &= 3.226(1.472K - 26.40) = 12.71\ . \end{aligned}\right\} \qquad \text{Eqs. (5)}$$

$$R^2 = .404 - \frac{5.36 \times 6.08}{23 + 102.32} = .144. \qquad \text{Eqs. (6) \& (11)}$$

For condition II we have

$$L = \frac{23 + 102.75}{5.70} = 22.061. \qquad \text{Eqs. (7a) \& (11)}$$

$$\left.\begin{aligned} b_1 &= .812\ (\ .122L + 3.30) = 4.86 \\ b_2 &= 2.191\ (\ .562L - 9.13) = 7.16 \\ b_3 &= 2.966\ (1.472L - 28.76) = 11.02. \end{aligned}\right\} \qquad \text{Eqs. (7)}$$

$$R^2 = .404 - \frac{5.36 \times 6.08}{23 + 102.32} = .146. \qquad \text{Eqs. (8) \& (11)}$$

For condition III we have

$$G = \frac{87.396 \times 120.361 - (148.322)^2}{2[6.074 \times 148.322 - 120.361 \times 5.362]} = 22.463.$$

Eqs. (9a) & (11)

$$K = \frac{87.396 \times 6.074 - 5.362 \times 148.322}{2[6.074 \times 148.322 - 120.361 \times 5.362]} = -.5175.$$

Eqs. (10) & (11)

$$\left.\begin{aligned} b_1 &= .612\ [\ .122G - \tfrac{1}{2}(-2.603) - K(-3.885)] = 3.70 \\ b_2 &= 2.222\ [\ .562G - \tfrac{1}{2}(\ 8.392) - K(\ 9.811)] = 7.44 \\ b_3 &= 3.226\ [1.472G - \tfrac{1}{2}(26.403) - K(31.254)] = 11.85\ . \end{aligned}\right\} \qquad \text{Eqs. (9)}$$

$$R^2 = \frac{\left[\begin{aligned}&(.404 \times 87.369 \times 120.361) + (2 \times 5.362 \times 6.074 \times 148.322) \\ &- .404(148.322)^2 - 87.396 \times (6.074)^2 - 120.361 \times (5.362)^2\end{aligned}\right]}{87.396 \times 120.361 - (148.322)^2} = .145.$$

Eqs. (12) & (11)

It will be noted that for all three conditions the rank order of the optimal times is the same. The R^2's for all three differ only in

the third decimal place. As a matter of fact, the optimal test lengths for the three conditions do not vary greatly from the original of 3, 10, 10, which means that the original lengths were not far from optimal. This accounts for the fact that the original R^2 of .144 is close to the other R^2's.

It should be noted that although the three conditions give nearly the same values for R^2, condition I is lowest of the three and condition II highest. This is according to theory.

Both cases I and III would be especially appropriate in the large scale use of IBM scoring machines by numerous operators working under unknown conditions of supervision. Case III gives the theoretically correct solution. It is probable that in many cases, however, as in the above example, the differences in validity for all three cases may not be significant. The R's and the b's for cases I and II are considerably easier to calculate than for case III. Case II can not be less than either I or II, hence may be regarded as an upper limit for case III. Case I would be a lower limit. In general, then, if one were determining optimal lengths for a battery of unit weights, he would calculate the R's for cases I and II. If they did not differ significantly, he would calculate the b's for case I, knowing that the results would not be significantly poorer than for case III, which requires considerably more computation.

If, however, the R for case II should be significantly higher than that for case I, it would probably be best to use equations (9) for calculating the test lengths.

If the problem of computing weighted composite scores is not regarded as crucial then, in general, the case II method should be used and test lengths computed by equations (7).

Appendix: Proof of the Solution for Condition III

We let

D_a = a diagonal matrix whose elements are the lengths of the original tests,

D_b = a diagonal matrix whose elements are the lengths of the altered test,

D_{σ_a} = a diagonal matrix of the standard deviations of the original tests,

P_a = the matrix covariances of the original tests,

P_{ac} = the column vector of validity covariances of the original tests,

P_b = the matrix of covariances of the altered tests,
P_{bc} = the column vector of validity covariances of the altered tests,
D_u = a diagonal matrix whose elements are one less the test reliabilities, and
R_b = the correlation between a criterion and a score on a battery of tests all of which have unit weight.

It can readily be shown that

$$R_b^2 = \frac{1' P_{bc} P'_{bc} 1}{1' P_b 1}, \tag{1}$$

where 1 is a column vector all of whose elements are unity. It can also be proved that

$$P_b = D_b D_a^{-1} P_a D_a^{-1} D_b - D_b^2 D_a^{-1} D + D_b D , \tag{2}$$

$$P_{bc} = D_b D_a^{-1} P_{ac} , \tag{3}$$

where

$$D = D_u D_a^{-1} D_{\sigma_a^2}. \tag{4}$$

Substituting (4) in (2), we have

$$P_b = D_b D_a^{-1} P_a D_a^{-1} D_b - D_b D_a^{-1} D_{\sigma_a} D_u D_{\sigma_a} D_a^{-1} D_b + D_b D_a^{-1} D_u D_{\sigma_a^2}. \tag{5}$$

Writing (1) for the tests of altered length from (3) and (5) yields

$$R_b^2 = \frac{1' D_b D_a^{-1} P_{ac} P_{ac}' D_a^{-1} D_b 1}{1'[D_b D_a^{-1} P_a D_a^{-1} D_b - D_b D_a^{-1} D_{\sigma_a} D_u D_{\sigma_a} D_a^{-1} D_b + D_b D_a^{-1} D_u D_{\sigma_a^2}]1}. \tag{6}$$

We write $D_b 1 = V_b$,

$$P_{ac} = D_{\sigma_a} r_{ac}, \tag{7}$$

$$P_a = D_{\sigma_a} r_a D_{\sigma_a}, \tag{8}$$

and substitute (7) and (8) in (6) to get

$$R_b^2 = \frac{V'_b D_a^{-1} D_{\sigma_a} r_{ac} r'_{ac} D_{\sigma_a} D_a^{-1} V_b}{V'_b D_a^{-1} D_{\sigma_a}(r_a - D_u) D_{\sigma_a} D_a^{-1} V_b + V'_b D_u D_{\sigma_a}{}^2 D_a^{-1} 1}. \tag{9}$$

Dropping the second subscript on D_{σ_a} and dropping the subscript of r_a , we let

$$\gamma = D_\sigma D_a^{-1} V_b . \tag{10}$$

We substitute (10) in (9) and get

$$R_b^2 = \frac{\gamma' r_{ac} r'_{ac} \gamma}{\gamma' (r - D_u)\gamma + \gamma' D_u D_\sigma 1}. \tag{11}$$

From (10),

$$D_\sigma^{-1} D_a \gamma = V_b. \tag{12}$$

Now we wish to determine γ so as to maximize (11) with the condition that the sum of the b's is some specified value. We write (11)

$$R_b^2 = \frac{f_1}{f_2}, \tag{13}$$

$$1' V_b = 1' D_\sigma^{-1} D_a \gamma = f_3. \tag{14}$$

From (13) and (14) we write

$$\phi = \frac{f_1}{f_2} + \lambda f_3, \tag{15}$$

where λ is the Lagrangian multiplier. Taking differentials of both sides of (15), we find

$$d\phi = \frac{1}{f_2} d f_1 - \frac{f_1}{f_2} 2\, d f_2 + \lambda\, d f_3. \tag{16}$$

From (11), (13), (14), and (16), we get

$$\frac{f_2}{2}\left[\frac{\partial \phi}{\partial \gamma'} + \left(\frac{\partial \phi}{\partial \gamma}\right)'\right] = 0 =$$

$$2\left\{r_{ac} f_1^1 - \frac{f_1}{f_2}\left[(r - D_u)\gamma + \tfrac{1}{2} D_u D_\sigma 1\right] + \frac{f_2}{2} \lambda D_a D_\sigma^{-1} 1\right\}. \tag{17}$$

We let

$$r_{ac} = U_1 \tag{18}$$

$$D_\sigma D_u 1 = U_2 \tag{19}$$

$$D_\sigma^{-1} D_a 1 = U_3 \tag{20}$$

$$(r - D_u) = \rho. \tag{21}$$

Substituting equations (18) through (21) in (17), equating to zero, and solving for γ, we get:

$$\gamma = \rho^{-1}(GU_1 - \tfrac{1}{2}U_2 + KU_3) \tag{22}$$

where

$$G = \frac{f_2}{f_1^{\frac{1}{2}}}, \quad \text{and} \tag{23}$$

$$K = \frac{f_2^2 \lambda}{2f_1}. \tag{24}$$

From (12) and (22) we have

$$V_b = D_\sigma^{-1} D_a \rho^{-1}(GU_1 - \tfrac{1}{2}U_2 + KU_3), \tag{25}$$

which gives the formal solution for the b's. However, we still have the unknowns G and K to solve for. We let

$$U'_i \rho^{-1} U_j = U'_j \rho^{-1} U_i = A_{ij}. \tag{26}$$

From (11), (13), and (18),

$$f_1^{\frac{1}{2}} = U'_1 \gamma. \tag{27}$$

From (11), (13) and (19), and (21),

$$f_2 = \gamma' \rho \gamma + U'_2 \gamma. \tag{28}$$

From (14) and (20),

$$1' V_b = U'_3 \gamma = T. \tag{29}$$

From (22), (26), and (27),

$$f_1^{\frac{1}{2}} = GA_{11} - \tfrac{1}{2}A_{12} + KA_{13}. \tag{30}$$

From (22), (26), and (28),

$$\begin{aligned} f_2 = {} & G(GA_{11} - \tfrac{1}{2}A_{12} + KA_{13}) + \tfrac{1}{2}(GA_{12} - \tfrac{1}{2}A_{22} + KA_{23}) \\ & + K(GA_{13} - \tfrac{1}{2}A_{23} + KA_{33}). \end{aligned} \tag{31}$$

From (22), (26), and (29),

$$T = GA_{13} - \tfrac{1}{2}A_{23} + KA_{33}. \tag{32}$$

Now let

$$E_i = GA_{i1} - \tfrac{1}{2}A_{i2} + KA_{i3}. \tag{33}$$

Substituting (33) in (30), (31), and (32) respectively, we have

$$f_1^{\frac{1}{2}} = E_1, \tag{34}$$

$$f_2 = GE_1 + \tfrac{1}{2}E_2 + KE_3, \tag{35}$$

$$T = E_3 . \tag{36}$$

From (23) and (34),

$$Gf_1^{\frac{1}{2}} = f_2 = GE_1 . \tag{37}$$

From (36),

$$KT = KE_3 . \tag{38}$$

Subtracting (37) and (38) from (35) yields

$$-2KT = E_2 . \tag{39}$$

But from equation (33), we can write (37), (39), and (38) respectively

$$\left.\begin{aligned} f_1^{\frac{1}{2}} &= GA_{11} - \tfrac{1}{2}A_{12} + KA_{13} , \\ -2KT &= GA_{12} - \tfrac{1}{2}A_{22} + KA_{23} , \\ T &= GA_{13} - \tfrac{1}{2}A_{23} + KA_{33} . \end{aligned}\right\} \tag{40}$$

Now let

$$H_1 = \frac{G}{f_1^{\frac{1}{2}}}, \tag{41}$$

$$H_2 = -\frac{1}{2f_1^{\frac{1}{2}}}, \tag{42}$$

$$H_3 = \frac{K}{f_1^{\frac{1}{2}}}. \tag{43}$$

With the aid of equations (41), (42), and (43), we rewrite (40)

$$\left.\begin{aligned} 1 &= H_1 A_{11} + H_2 A_{12} + H_3 A_{13} \\ 0 &= H_1 A_{12} + H_2 A_{22} + H_3 (A_{23} + 2T) \\ 0 &= H_1 A_{13} + H_2 (A_{23} + 2T) + H_3 A_{33} . \end{aligned}\right\} \tag{44}$$

If we let

H = a vector of the H_i's,
U = a matrix of the U_i's, i.e., (U_1 , U_2 , U_3),
e_i = a vector all of whose elements are zero but the ith, which is 1, and
e_{ij} = a 3×3 matrix all of whose elements are zero but the ijth, which is 1,

then (44) can be rewritten in matrix notation

$$e_1 = [U' \rho^{-1} U + 2T(e_{23} + e_{32})] H \tag{45}$$

or

$$H = [U' \rho^{-1} U + 2T(e_{23} + e_{32})]^{-1} e_1, \tag{46}$$

which gives the solution for the H_i's. Then from (13), (23), and (41)

$$R_b^2 = \frac{1}{H_1}. \tag{47}$$

From (41) and (42),

$$\frac{-H_1}{2H_2} = G. \tag{48}$$

From (42) and (43),

$$\frac{-H_3}{2H_2} = K. \tag{49}$$

It can be shown that

$$A_{11} = R_m^2, \tag{50}$$

where R_m^2 is the maximum possible correlation assuming all tests of infinite length or perfect reliability.

The values given by (47), (48), and (49) can be readily solved for in terms of T and the A_{ij} from equations (44). Using also (50), they are respectively

$$R^2 = \frac{\left[\begin{array}{l} R_m^2 A_{22} A_{33} + 2A_{12} A_{13}(A_{23} + 2T) \\ - [R_m^2 (A_{23} + 2T)^2 + A_{22} A_{13}^2 + A_{33} A_{12}^2] \end{array}\right]}{A_{22} A_{33} - (A_{23} + 2T)^2}, \tag{51}$$

$$G = \frac{(2T + A_{23})^2 - A_{22} A_{33}}{2\,[A_{13}(2T + A_{23}) - A_{33} A_{12}]}, \tag{52}$$

$$K = \frac{A_{22} A_{13} - A_{12}(2T + A_{23})}{2\,[A_{13}(2T + A_{23}) - A_{33} A_{12}]}. \tag{53}$$

REFERENCES

1. Horst, Paul. Regression weights as a function of test length. *Psychometrika*, 1948, 13, 125-132.
2. Horst, Paul. Determination of optimal test length to maximize the multiple correlation. *Psychometrika* 1949, 14, 79-88.

Manuscript received 5/1/50.
Revised manuscript received 6/7/50.

PSYCHOMETRIKA—VOL. 16, NO. 2
JUNE, 1951

REMARKS ON THE METHOD OF PAIRED COMPARISONS: II. THE EFFECT OF AN ABERRANT STANDARD DEVIATION WHEN EQUAL STANDARD DEVIATIONS AND EQUAL CORRELATIONS ARE ASSUMED*

FREDERICK MOSTELLER
HARVARD UNIVERSITY

If customary methods of solution are used on the method of paired comparisons for Thurstone's Case V (assuming equal standard deviations of sensations for each stimulus), when in fact one or more of the standard deviations is aberrant, all stimuli will be properly spaced except the one with the aberrant standard deviation. A formula is given to show the amount of error due to the aberrant stimulus.

1. *Introduction.* In a previous article† we showed that the ordinary solution to Thurstone's Case V of the method of paired comparisons was a least-squares solution. It was also pointed out that for Case V it was not necessary to assume that all correlations between stimulus sensations were zero; it was sufficient to assume the correlations were equal. Thurstone's Case V assumes that all standard deviations of stimulus sensations are equal. In this article we will investigate the effect of an aberrant standard deviation on the Case V solution. We will deal with error-free data.

2. *The Problem of the Aberrant Standard Deviation.* As in the previous article, we suppose the objects $0_1, 0_2, \cdots, 0_n$ to have sensation means $S_1, S_2, \cdots, S_n$. We shall also assume

$$\begin{aligned} &\text{the standard deviation of } X_i = \sigma_i = \sigma, \quad (i = 1, 2, \cdots, n-1); \\ &\text{the standard deviation of } X_n = \sigma_n; \text{ and} \\ &\text{the correlation } \rho_{ij} = \rho. \end{aligned} \tag{1}$$

In other words, the standard deviations are all equal except the one associated with S_n, and the correlations are all equal.

*This research was performed in the Laboratory of Social Relations under a grant made available to Harvard University by the RAND Corporation under the Department of the Air Force, Project RAND.

†Mosteller, Frederick. Remarks on the Method of Paired Comparisons: I. The Least Squares Solution Assuming Equal Standard Deviations and Equal Correlations. *Psychometrika*, 1951, **16**, 3-9.

Then we may define the matrix of differences between means in original standard deviation units as

$$X_{ij} = \frac{S_i - S_j}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho\sigma_i\sigma_j}}, \quad (i,j = 1,2,\cdots,n); \tag{2}$$

or

$$X_{ij} = \frac{S_i - S_j}{\sqrt{2\sigma^2(1-\rho)}}, \quad (i,j \neq n).$$
$$X_{in} = \frac{S_i - S_n}{\sqrt{\sigma^2 + \sigma_n^2 - 2\rho\sigma\sigma_n}}. \tag{3}$$

There is no loss of generality and great gain in convenience if we define

$$2\sigma^2(1-\rho) = 1. \tag{4}$$

Also

$$\sigma^2 + \sigma_n^2 - 2\rho\sigma\sigma_n = \sigma_d^2. \tag{5}$$

Now if we write our X_{ij} matrix we have:

X_{ij} MATRIX

	1	2	3	.....	n
1	$S_1 - S_1$	$S_1 - S_2$	$S_1 - S_3$		$(S_1 - S_n)/\sigma_d$
2	$S_2 - S_1$	$S_2 - S_2$	$S_2 - S_3$		$(S_2 - S_n)/\sigma_d$
3	$S_3 - S_1$	$S_3 - S_2$	$S_3 - S_3$		$(S_3 - S_n)/\sigma_d$
.					
.					
.					
n	$(S_n - S_1)/\sigma_d$	$(S_n - S_2)/\sigma_d$	$(S_n - S_3)/\sigma_d$		$(S_n - S_n)/\sigma_d$

We will work out the least-squares solution much as described in the earlier article as if σ_d were unity. That is, we behave as if the standard deviations are equal as we would if we were experimenters using Case V. This merely involves summing the columns and averaging.

From this matrix the total for the ith column is

$$S_i^* = \sum_{j=1}^{n-1} S_j - (n-1)S_i + (S_n - S_i)/\sigma_d, \quad (i = 1,2,\cdots,n-1); \tag{6}$$

$$S_n^* = \left(\sum_{i=1}^{n} S_j - nS_n\right)\sigma_d, \qquad (i = n).$$

The S_i^* are essentially estimates of the least-squares solution when the standard deviations are in fact equal. We can perform linear transformations on them without changing the symbol. Because the result would only be good to a linear transformation, we are allowed to subtract $\sum_{j=1}^{n-1} S_j$ from all these results, and then we temporarily set $S_n = 0$. This gives

$$\begin{aligned} S_i^* &= -S_i[n-1+(1/\sigma_d)], \qquad (i=1,2,\cdots,n-1); \\ S_n^* &= \sum_{j=1}^{n-1} S_j[(1/\sigma_d)-1]. \end{aligned} \tag{7}$$

We may change the scale factor assumed in equation (4) by multiplying through by

$$\frac{-1}{(n-1)+1/\sigma_d},$$

and this at last gives

$$\begin{aligned} S_i^* &= S_i \qquad\qquad (i=1,2,\cdots,n-1), \\ S_n^* &= [(1-1/\sigma_d)/(n-1+1/\sigma_d)] \sum_{j=1}^{n-1} S_j; \end{aligned} \tag{8}$$

or, since $S_n = 0$,

$$S_n^* = [n(1-1/\sigma_d)/(n-1+1/\sigma_d)]\,\bar{S}, \qquad (S_n=0).$$

The gratifying part of this result is that all the S_i are properly spaced relative to one another except S_n. In other words, changing one of the standard deviations affects only the position of the object with the aberrant sensation standard deviation. We note, of course, that when $\sigma_d^2 = 1$, [i.e., $= 2\sigma^2(1-\rho)$], $S_n^* = 0$ as anticipated. We also note that when the grand mean $\bar{S}$ is small, that is when S_n is centrally located with respect to the other stimuli means, the effect of an aberrant stimulus is small. Thus if we have reason to believe that some particular object has a much different sensation variability from the rest, the other objects should be so chosen that the aberrant one is near the center of the scale, or else it should be excluded.

If we suppose $\sigma_d > 1$, and n of reasonable size, we may approximate S_n^* by

$$S_n^* = (1 - 1/\sigma_d)\,\bar{S}\,.$$

3. *Examples.*

(a) Suppose the values $S_1, S_2, \cdots, S_6$ are $-4, -2, 0, 1, 2, 3$ and S_3 has a standard deviation different from the rest. Application of equation (9) shows that the spacing will be correct.

(b) With the same S values as in Example (a), suppose S_1 has $\sigma_d = 2$. Then $S_2, \cdots, S_6$ will be properly scaled with values $2, 4, 5, 7, 10$ (we must add 4 to all values because we take $S_1 = 0$) and

$$S_n^* = \frac{6(1 - \frac{1}{2})}{5 + \frac{1}{2}} \frac{28}{6} = 2.55$$

instead of zero.

(c). With the same S values as in Example (a), suppose S_1 has $\sigma_d = 1/2$. Then $S_2, \cdots, S_6$ will again be properly scaled as in Example (b), but $S_n^* = -4$ instead of zero.

4. *Generalization to Several Aberrant Standard Deviations.* Although it will not be shown here, the generalization to several aberrant standard deviations is immediate. If we have a set of objects $0_1, 0_2, \cdots, 0_n$, with variances $\sigma^2, \sigma^2, \sigma^2, \cdots, \sigma_{n-k}^2, \sigma_{n-k+1}^2, \cdots, \sigma_n^2$, then the standard method of solving paired comparisons, Case V, will leave those stimuli with equal variances appropriately spaced. Of course, there need to be at least three stimuli with equal variances for this result to be interesting or useful.

It follows that if we have two or more sets of stimuli such that the standard deviations within each set are equal, each set will itself be properly spaced, but the sets will not be spaced or positioned correctly relative to one another.

It is conceivable that in a practical situation a different method could be used for some of the measurements, so that we could get an estimate of the relative sizes of the sigmas and that this information could be useful in practice.

Thurstone has already noticed that small changes in the sigmas do not affect the solution much.

Manuscript received 8/22/50.

PSYCHOMETRIKA—VOL. 16, NO. 2
JUNE, 1951

REMARKS ON THE METHOD OF PAIRED COMPARISONS: III. A TEST OF SIGNIFICANCE FOR PAIRED COMPARISONS WHEN EQUAL STANDARD DEVIATIONS AND EQUAL CORRELATIONS ARE ASSUMED*

FREDERICK MOSTELLER
HARVARD UNIVERSITY

A test of goodness of fit is developed for Thurstone's method of paired comparisons, Case V. The test involves the computation of

$$\chi^2 = n\Sigma(\theta'' - \theta')^2/821,$$

where n is the number of observations per pair, and θ'' and θ' are the angles obtained by applying the inverse sine transformation to the fitted and the observed proportions respectively. The number of degrees of freedom is $(k-1)(k-2)/2$.

1. *Introduction*

It would be useful in Thurstone's method of paired comparisons to have a measure of the goodness of fit of the estimated proportions to the observed proportions. Ideally we might try to find estimates of the stimuli positions S_i such that we can reproduce the observed proportions p'_{ij} as closely as possible in some sense.

One kind of test might be based on

$$\chi^2 = \frac{\Sigma(p''_{ij} - p'_{ij})^2}{\sigma^2_{ij}}$$

where p''_{ij} is the estimate of p'_{ij} derived from the S'_i. But the true p_{ij} are not known and would have to be replaced by the observed p'_{ij}. If one does replace the p_{ij} by p'_{ij} and σ_{ij} by σ'_{ij}, then it is possible to fit the S'_i by means of a minimum chi-square criterion. However, such a procedure calls for an iterative scheme and involves extremely tedious computations. An alternative method is suggested by the inverse sine transformation.

*This research was performed in the Laboratory of Social Relations under a grant made available to Harvard University by the RAND Corporation under the Department of the Air Force, Project RAND.

2. *The model*

It is assumed that we have a set of stimuli which, when presented to a subject, produce sensations. These sensations are assumed to be normally distributed, perhaps with different means. However the standard deviations of each distribution are assumed to be the same, and the correlations between pairs of stimuli sensations are assumed equal.

Subjects are presented with pairs of stimuli and asked to state which member of each pair is greater with respect to some property attributed to all the stimuli (the property is the dimension of the scale we are trying to form). Our observations consist of the proportions of times stimulus j is judged "greater than" stimulus i. We call these proportions p'_{ij} to indicate that they are observations and not the true proportions p_{ij}.

From the observed proportions we compute normal deviates X'_{ij} and proceed in the usual way (5) to estimate the stimulus positions, S'_i, on the sensation scale. Once the S'_i are found we can retrace to get the fitted normal deviates X''_{ij} and the fitted proportions p''_{ij}.

Our problem is to provide a method for ascertaining how well the fitted p''_{ij} agree with the observed p'_{ij}.

In such a test of significance involving goodness of fit, we are interested in knowing what the null hypothesis and the alternative hypothesis are. In the present case the null hypothesis is given by the model assumed above. However, the alternative hypothesis is quite general: merely that the null hypothesis is not correct. In particular, the null hypothesis assumes additivity so that if D_{ij} is the distance from S_i to S_j and D_{jk} is the distance from S_j to S_k, we should find

$$D_{ik} = D_{ij} + D_{jk}.$$

If we do not have unidimensionality this additivity property will usually not hold.

For example, consider the case of three stimuli with $S_1 < S_2 < S_3$. If the standard deviation of each distribution is the same, we might write

$$\begin{aligned} D_{12} &= S_2 - S_1 \\ D_{13} &= S_3 - S_1 \\ D_{23} &= S_3 - S_2. \end{aligned}$$

Since we can choose $S_1 = 0$ and $S_2 = D_{12}$, S_3 from the second equation must be D_{13}. Finally

$$D_{23} = D_{13} - D_{12}.$$

Since each of our comparisons of stimuli is done independently it is not necessary that this relation hold either for the observations or for the theoretical values. Indeed the observed value of D_{23} could have conflicted with the assumption of additivity. Such a failure of additivity makes the fitting of the observed p'_{ij} less likely, and on the average failure will increase the value of χ^2 in our test.

It can also happen that the standard deviations of the various stimuli are not equal even though unidimensionality obtains. In this case our attempt to fit the data under the equal standard deviations assumption will sometimes fail, and this failure will be reflected, in general, in a failure of additivity and thus an increase in χ^2.

3. *The transformation*

Like so many other good things in statistics, the inverse sine transformation was developed by R. A. Fisher (4). Further discussion by Bartlett (1, 2), Eisenhart (8), and Mosteller and Tukey (7) may be of interest to those who wish to examine the literature. The facts essential to the present discussion are these: If we have an observed p' arising from a binomial sample of size n from a population with true proportion of successes p, then

$$\theta' = \text{arc sin } \sqrt{p'} \tag{1}$$

is approximately normally distributed with variance

$$\sigma_{\theta'}^2 = \frac{821}{n}, \tag{2}$$

nearly independent of the true p, when θ' is measured in degrees. A table for making the transformation to angles has been computed by C. I. Bliss (3), and is readily available in G. W. Snedecor's *Statistical Methods* (4th Edition), p. 450.

Then if we define

$$\begin{aligned} \theta'_{ij} &= \text{arc sin } \sqrt{p'_{ij}} \\ \theta''_{ij} &= \text{arc sin } \sqrt{p''_{ij}} \end{aligned} \tag{3}$$

where p'_{ij} are the observed proportions and p''_{ij} are the proportions derived from fitting the S_i, we can test goodness of fit by

$$\chi^2 = \sum_{i<j} \frac{(\theta''_{ij} - \theta'_{ij})^2}{821/n}. \tag{4}$$

If there are k stimuli we have k parameters to fit, the k S'_i values. But two of these are the zero point and the scale factor, which are arbitrary. This leaves k—2 parameters free for fitting the data. There are $k(k\text{—}1)/2$ p'_{ij}'s to be fitted. So it appears that the appropriate number of degrees of freedom for the test is $k(k\text{—}1)/2 - (k\text{—}2) - 1 = (k\text{—}1)(k\text{—}2)/2$. We note that with two stimuli we can always fit the data perfectly, so there should be zero degrees of freedom as the formula indicates.

4. *Illustrative example*

To illustrate the test we will use the paired comparison method on the American League baseball record for 1948. The following table gives the observed p'_{ij}. The number in the ith row and jth column is the proportion of games won by the team named at the top of the jth column from the team named at the left of the ith row. In this situation we regard the clubs as stimuli which have distributions of performances. The number of games each club plays with each other club is 22 (except for minor fluctuations). Successive tables indicate the steps in the solution. The steps are these:

1. From p'_{ij} table obtain X'_{ij} table from a table of the normal integral.
2. Solve for the S'_i by summing columns and averaging.
3. Use S'_i to obtain X''_{ij}, $X''_{ij} = S'_i - S'_j$.
4. Use X''_{ij} to obtain p''_{ij}, from a table of the normal integral.
5. Compute θ'', θ', $\theta'' - \theta'$.
6. Get the sum of squares of $\theta'' - \theta'$.
7. Divide the sum of squares by $821/n$, here $821/22$.
8. Look up result in χ^2 table with $(k\text{—}1)(k\text{—}2)/2$ degrees of freedom.

PROPORTIONS OF ALL GAMES THAT THE TEAM GIVEN AT THE TOP OF THE COLUMN WON FROM THOSE AT THE LEFT (1948)
Each Entry Represents 22 Games

p'_{ij} Table

	Clev.	Bost.	N.Y.	Phil.	Det.	St.L.	Wash.	Chic.
Clev.	—	.478	.545	.273	.409	.364	.273	.273
Bost.	.522	—	.364	.455	.318	.318	.318	.364
N.Y.	.455	.636	—	.455	.409	.273	.227	.273
Phil.	.727	.545	.545	—	.545	.182	.364	.273
Det.	.591	.682	.591	.455	—	.500	.273	.364
St.L.	.636	.682	.727	.818	.500	—	.545	.381
Wash.	.727	.682	.773	.636	.727	.455	—	.429
Chic.	.727	.636	.727	.727	.636	.619	.571	—

X'_{ij} Table

	Clev.	Bost.	N.Y.	Phil.	Det.	St.L.	Wash.	Chic.
Clev.	——	—.055	+.113	—.604	—.230	—.348	—.604	—.604
Bost.	+.055	——	—.348	—.113	—.473	—.473	—.473	—.348
N.Y.	—.113	+.348	——	—.113	—.230	—.604	—.749	—.604
Phil.	+.604	+.113	+.113	——	+.113	—.908	—.348	—.604
Det.	+.230	+.473	+.230	—.113	——	.000	—.604	—.348
St.L.	+.348	+.473	+.604	+.908	.000	——	+.113	—.303
Wash.	+.604	+.473	+.749	+.348	+.604	—.113	——	—.179
Chic.	+.604	+.348	+.604	+.604	+.348	+.303	+.179	——
	+2.332	+2.173	+2.065	+0.917	+0.132	—2.143	—2.486	—2.990
S'_i	.2915	.2716	.2581	.1146	.0165	—.2678	—.3108	—.3738

X''_{ij} Table

	Clev.	Bost.	N.Y.	Phil.	Det.	St.L.	Wash.	Chic.
Clev.	——							
Bost.	.0199	——						
N.Y.	.0334	.0135	——					
Phil.	.1769	.1570	.1435	——				
Det.	.2750	.2551	.2416	.0981	——			
St.L.	.5593	.5394	.5259	.3824	.2843	——		
Wash.	.6023	.5824	.5689	.4254	.3273	.0430	——	
Chic.	.6653	.6454	.6319	.4884	.3903	.1060	.0630	——

p''_{ij} Table

	Clev.	Bost.	N.Y.	Phil.	Det.	St.L.	Wash.	Chic.
Clev.	——							
Bost.	.508	——						
N.Y.	.513	.505	——					
Phil.	.570	.562	.557	——				
Det.	.608	.601	.595	.539	——			
St.L.	.712	.705	.700	.649	.612	——		
Wash.	.726	.720	.715	.665	.628	.517	——	
Chic.	.747	.741	.736	.687	.652	.542	.525	——

Table of θ'', θ', $\theta'' - \theta'$

	Clev.	Bost.	N.Y.	Phil.	Det.	St.L.	Wash.	Chic.
Clev.								
	45.46							
Bost.	46.26							
	—.80							
	45.75	45.29						
N.Y.	42.42	52.89						
	+3.33	—.760						
	49.02	48.56	48.27					
Phil.	58.50	47.58	47.58					
	—9.48	+0.98	+0.69					
	51.24	50.83	50.48	47.24				
Det.	50.24	55.67	50.24	42.42				
	+1.00	—4.84	+0.24	+4.82				
	57.54	57.10	56.79	53.67	51.47			
St.L.	52.89	55.67	58.50	64.75	45.00			
	+4.65	+1.43	—1.71	—11.08	+6.47			
	58.44	58.05	57.73	54.63	52.42	45.97		
Wash.	58.50	55.67	61.55	52.89	58.50	42.42		
	—.06	+2.38	—3.82	+1.74	—6.08	+3.55		
	59.80	59.41	59.08	55.98	53.85	47.41	46.43	
Chic.	58.50	52.89	58.50	58.50	52.89	51.88	49.08	
	+1.30	+6.52	+0.58	—2.52	+0.96	—4.47	—2.65	

$$\Sigma(\theta'' - \theta')^2 = 551.40$$
$$821/22 = 37.32$$

$$\chi^2_{21} = 14.78 \qquad .80 < P(\chi^2) < .90$$

The chi-square result shows rather good agreement between the fitted data and the observed data. Investigation of additional baseball data has suggested that the agreement is usually too good rather than not good enough. It was suggested to the author that a possible reason for this is that the proportion of games won by any team from another team involves an admixture of games played at home and away, and that if these were separated we might then not get such consistently good agreement. As an example, suppose probabilities

of winning at home and away are .25 and .75 respectively, averaging .50. The variance of games won based on the $p = .50$ is $n/4$, but based on $n/2$ games at .25 and $n/2$ at .75, the variance is $3n/16$, somewhat smaller. The decrease in variance would be similar to that gained from stratified sampling. Calculations not presented here suggest that this may be the case.

It should be remembered that we have found the best S'_i's in the least-squares sense to reproduce the X'_{ij}'s, and have not done our best to reproduce the θ's. This means that, had we done a more elaborate method of fitting, we might have obtained a still better fit and consequently a higher value of P (which is already quite high).

5. *The power of the test for three stimuli*

The power of the test developed, that is the probability of rejecting the null hypothesis when it is false, is rather awkward to investigate. The power depends on the degree of divergence from the assumptions, the number of stimuli involved, the number of observations for each pair of stimuli, as well as the significance level chosen. We will discuss the power for a rather special case. This case has the advantage that it displays the workings of the chi-square test rather clearly and is easy to compute. Our procedure will be: (1) set up the model, (2) compute χ^2 for this case, (3) insert a departure from the model, (4) investigate the power for the special case under consideration.

We will assume that the standard deviations of the differences between pairs of stimuli are unity. The true stimuli means are in the order $S_3 > S_2 > S_1$. Furthermore we will assume that these means are sufficiently close to one another that the approximation

$$p_{ij} = \frac{1}{\sqrt{2\pi}} \int_{-(S_j - S_i)}^{\infty} e^{-\frac{1}{2}x^2}\, dx \cong \frac{1}{2} + \frac{S_j - S_i}{\sqrt{2\pi}}, \tag{5}$$

will be adequate. For this case p_{ij} will be nearly $1/2$, so we will be able to use the approximation:

$$\sigma^2(p'_{ij}) = \frac{1}{4n} = \sigma^2. \tag{6}$$

Working with this case will have the further advantage that we will not need to use the inverse sine transformation but can work directly with

$$\chi^2 = \frac{\sum_{i<j} (p'_{ij} - p''_{ij})^2}{\sigma^2}, \tag{7}$$

since our principal reason for working with the transformation was that σ^2 was not known.

The observations can be written

$$p'_{ij} = p_{ij} + k_{ij}\sigma . \tag{8}$$

Here the unprimed p is the true proportion of the time stimulus j is reported to exceed stimulus i, the primed p is the corresponding observed proportion, σ is $1/4n$, and k_{ij} is a random normal deviate with zero mean and standard deviation unity. The sample size is n assumed to be reasonably large.

Under these assumptions

$$\begin{aligned} p'_{ij} = p_{ij} + k_{ij}\sigma &= \frac{1}{\sqrt{2\pi}} \int_{-(S_j - S_i) + \xi_{ij}}^{\infty} e^{-\frac{1}{2}x^2}\, dx \cong \frac{1}{2} + \frac{S_j - S_i - \xi_{ij}}{\sqrt{2\pi}} \\ &\cong \frac{1}{2} + \frac{S_j - S_i + k_{ij}\sigma\sqrt{2\pi}}{\sqrt{2\pi}}. \end{aligned} \tag{9}$$

Thus the normal deviate corresponding to p'_{ij} is approximately

$$D'_{ij} = S_j - S_i + k_{ij}\sigma\sqrt{2\pi} . \tag{10}$$

Now we insert these values in the paired comparison table as usual and solve for the estimates of the stimuli positions S'_i by summing columns and averaging. After adding the mean of the true stimuli positions these estimates are:

$$\begin{aligned} S'_1 &= S_1 - (k_{12} + k_{13})\sigma\sqrt{2\pi}/3 , \\ S'_2 &= S_2 + (k_{12} - k_{23})\sigma\sqrt{2\pi}/3 , \\ S'_3 &= S_3 + (k_{13} + k_{23})\sigma\sqrt{2\pi}/3 . \end{aligned} \tag{11}$$

We take the differences of these pairs to get the fitted normal deviates, the D''_{ij}:

$$\begin{aligned} D''_{12} &= S_2 - S_1 + (2k_{12} + k_{13} - k_{23})\sigma\sqrt{2\pi}/3 , \\ D''_{13} &= S_3 - S_1 + (k_{12} + 2k_{13} + k_{23})\sigma\sqrt{2\pi}/3 , \\ D''_{23} &= S_3 - S_2 + (-k_{12} + k_{13} + 2k_{23})\sigma\sqrt{2\pi}/3 . \end{aligned} \tag{12}$$

Now the fitted proportions p''_{ij} are approximately

$$p''_{ij} = \frac{1}{2} + \frac{D''_{ij}}{\sqrt{2\pi}}. \tag{13}$$

When we take the differences $p'_{ij} - p''_{ij}$ we get

$$\begin{aligned} p'_{12} - p''_{12} &= (k_{12} - k_{13} + k_{23})\sigma/3\,, \\ p'_{13} - p''_{13} &= (-k_{12} + k_{13} - k_{23})\sigma/3\,, \\ p'_{23} - p''_{23} &= (k_{12} - k_{13} + k_{23})\sigma/3\,. \end{aligned} \tag{14}$$

Now immediate computation of χ^2 inserting the values from equations (14) into equation (7) is

$$\chi^2 = \left(\frac{k_{12} - k_{13} + k_{23}}{\sqrt{3}} \right)^2. \tag{15}$$

Since the k's are normally and independently distributed with zero means and unit variance, the quantity in parentheses is in turn a normal deviate with zero mean and unit variance, because the standard deviation of the sum in the numerator is $\sqrt{3}$. Of course, the square of such a normal deviate is distributed like χ^2 with one degree of freedom. In this special case then we have shown how the χ^2 test arises.

We have incidentally set up the machinery for examining the power of the test for our special case. Until now we have assumed that the p_{ij} were arranged to get consistency in the spacings between the true stimuli means. We now relax this condition. In particular let us suppose that the consistent p_{23} is replaced by $p_{23} + \Delta$ where Δ is an error due to the lack of unidimensionality of the stimuli we are considering. This means that p'_{23} will be replaced by $p'_{23} + \Delta$, which in turn means that k_{23} will be replaced by $k_{23} + \Delta/\sigma$. Now when we come to compute χ^2 with the null hypothesis not satisfied we get

$$\chi^{2*} = (\chi + \Delta/\sqrt{3}\,\sigma)^2. \tag{16}$$

Here χ is a normal deviate, the expression inside the parentheses on the right of equation (15). If we are working with a significance test at the 5% level we will reject the null hypothesis unless

$$-1.96 < \chi + \Delta/\sqrt{3}\,\sigma < 1.96\,.$$

The following table indicates very roughly how often we will reject the null hypothesis as $\Delta/\sqrt{3}\,\sigma$ takes various values.

$\Delta/\sqrt{3}\,\sigma$	Percent rejected
1	16%
1.96	50%
2	52%
3	84%

We say roughly because when Δ takes large values our approximations no longer hold very well. Nevertheless these values are indicative of the magnitudes.

Let us see how much error there must be in p_{23} to raise the rejection level to 16%. Suppose $n = 48$. Then

$$\frac{\Delta^2}{3\sigma^2} = 1$$

$$\Delta^2 = 3\sigma^2 = \frac{3}{4 \times 48} = \frac{1}{64}$$

$$\Delta = .125 .$$

Thus for samples as large as 48, p_{23} must deviate from the consistent value of approximately .5 by as much as .125 to raise the probability of rejection from 5% to 16%.

A short discussion of the kinds of alternatives that can exist in paired comparisons and the general behavior of this test against these may assist the reader. The principal ways the Case V assumptions can be violated are

(1) lack of normality,

(2) lack of unidimensionality,

(3) failure of the equal standard deviation of differences assumption.

Failure of normality is not important to the method of paired comparisons, as we shall show elsewhere. It is just as well then that the present test will be very poor at detecting deviations from normality. The normality assumption is more in the nature of a computational device than anything else.

Lack of unidimensionality will be reflected in the failure of distances between estimated stimuli positions to agree with the observed distances, and thus we will have high chi-square values. The

principal alternative of interest then, is one for which the test is sensitive.

Unfortunately it is also possible that we have unidimensionality without having equality of standard deviations of differences of pairs. The result of using Case V may be to give a large chi-square value when this happens. This is not uniformly true however. It is possible to have unequal standard deviations without detecting this fact in the Case V solution as has been shown elsewhere (6). In particular, if there is only one aberrant standard deviation, and if the stimulus mean for that stimulus is near the mean of all the stimulus positions, the chi-square test will not be likely to detect this failure of the model. The best that can be said is that sometimes such aberrations will cause high values of chi-square and sometimes not, depending on the nature of the case.

We might like to relax our conditions and not use Case V but try to use some other case. However, this requires a large number of stimuli. In the case of the assumption of independence between pairs of stimuli we still have for k stimuli a total of k means and k variances to choose. Two of these $2k$ values are merely scale and location parameters, so we have in all $2k - 2$ things that can be varied as against $k(k-1)/2$ cell entries. Thus we need at least 5 stimuli to begin to get degrees of freedom for testing. With a reasonable number of stimuli we could still test for unidimensionality in the face of unequal stimulus variabilities. When we come to the completely general case, allowing the correlation coefficients to vary as well, the problem is hopeless. We now have more degrees of freedom at our disposal than there are in the table. It seems reasonable then never to try to test for unidimensionality under a more general assumption than equal correlations and unequal variances for the stimuli.

6. *Conclusions*

A test of the assumptions underlying Thurstone's method of paired comparisons is developed and illustrated. The inner workings of the test and an indication of its power are provided for a special case involving three stimuli lying very close to one another. Although the method is developed and applied for Thurstone's Case V, it can be applied to any paired comparison case providing some degrees of freedom are left over after the process of estimating the spacings between the stimuli positions has been completed.

REFERENCES

1. Bartlett, M. S. The square root transformation in analysis of variance. *Supp. J. roy. stat. Soc.*, 1936, **3**, 68-78.
2. Bartlett, M. S. The use of transformations. *Biometrics Bull.*, 1947, **3**, 39-52.
3. Bliss, C. I. Plant protection, No. 12. Leningrad, 1937.
4. Fisher, R. A. On the dominance ratio. *Proc. roy. soc. Edinb.*, 1922, **42**, 321-341.
5. Mosteller, F. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 1951, **16**, 3-9.
6. Mosteller, F. Remarks on the method of paired comparisons: II. The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. *Psychometrika*, 1951, **16**, 203-206.
7. Mosteller, F, and Tukey, J. W. The uses and usefulness of binomial probability paper. *J. Amer. statist. Ass.*, 1949, **44**, 174-212.
8. Statistical Research Group, Columbia University. Selected techniques of statistical analysis. New York: McGraw-Hill Book Co., 1947.

Manuscript received 9/2/50
Revised manuscript received 11/13/50

PSYCHOMETRIKA—VOL. 16, NO. 2
JUNE, 1951

RATE OF ADDITION AS A FUNCTION OF DIFFICULTY AND AGE*

JAMES E. BIRREN AND JACK BOTWINICK

Rate of addition was studied as a function of difficulty as measured by problem length. The hypothesis was tested that the rate of addition would decline as a function of the logarithm of the number of addition operations per problem. The test material required the rapid addition of single columns of digits ranging from two to twenty-five digits in length. Rate of uncorrected addition declined as a power function of problem length and the rate of correct addition declined as an exponential function of length. Results indicated that subjects who varied in age and mental status could be differentiated according to the parameters defining the curves of addition rate as a function of length.

Introduction

This study was designed to investigate the effect of problem length upon the rate of addition. The objectives were to determine whether: (a) a psychophysical relation exists between the rate of addition and difficulty as measured by problem length and (b) increasing problem difficulty results in disproportionate reduction in rate of addition in a class of individuals known to have altered mental function, e.g., senile mental patients.

A group of early investigators (6, 11, 12) studied rate of addition as a function of fatigue, drug effects, and other conditions. This work suggested using addition rate as the measure of performance. By changing the number of digits per problem it was anticipated that difficulty could be varied systematically and rate of performance could be described as a function of the number of operations comprising the task.

The ratio of change in difficulty to change in performance would describe a continuous function if a uniform relation exists between these variables. This presupposes that there exists a rational basis for estimating the difficulty of a task and a satisfactory measure of performance; both are chronic issues in psychometric research. If

*The cooperation of the staff of the Baltimore City Schools, the cooperation of the staffs of the Spring Grove and the Springfield Maryland State Hospitals and the Sheppard and Enoch Pratt Hospital, the assistance of Miss Charlotte Fox in devising the test material and gathering some of the data, and the computational assistance of Mrs. Betty Benser are gratefully acknowledged.

a method of experimental control of difficulty were available, one could compare stimuli from diverse fields and accordingly, compare the subjects' abilities on an absolute basis (3, 4). One method of describing the magnitude of a stimulus is its length or number of composite elements. In the present study the measure of difficulty selected was problem length or the number of addition operations in a problem. If it is demonstrated that addition performance declines systematically as a function of problem length, then difficulty can be defined operationally by the number of addition operations.

In such a relation the investigator must select an appropriate measure of performance as well as of difficulty. An adequate measure of performance appears to lie in the use of time scores or rate of performance. Landahl (5) suggested the use of time scores in place of item scores and Thurstone (10) has also incorporated a time score in a conceptual relation between difficulty, time, and probability of completing the problem. Thurstone defined the ability of a subject as the difficulty level at which the probability is one-half that the subject will complete the problem in infinite time.

A distinct advantage of time scores is their applicability throughout the range of difficulty. Thus a time score gives information as to how well an individual handles a very simple item, i.e., how much time is required. In contrast a relative measure of difficulty, per cent passing, gives no discrimination among individuals for easy material. Coombs (2) has pointed out that one method of improving psychometric measurements would be to develop a method for collecting data which would enable us to know how well a subject passed an item or how badly he failed it. The use of time scores appears to be such a method and may be a key to developing more homogeneous tests. The use of the ratio of change in rate of performance to change in difficulty might yield "purer" psychological measurements than now obtainable since fewer constants would be involved. The investigator assumes with this approach that the subject's standard of accuracy or his level of motivation is unchanged during the test session (10). It is not necessary for such factors to be the same from person to person, only that they be constant within the individual. If constant, their effect would be removed from the derivative of the function, i.e., the change in rate of performance as a function of change in difficulty.

Tate (9) found evidence of a general speed factor as well as specific speed abilities in performance of several types of test material. The demonstration of specific and general speed factors offers

encouragement to the development of rational measures of difficulty. By the use of time scores, abilities could be analyzed in a difficulty range unmeasurable by item scores.

Methods

Materials: Pages of addition problems consisting of single columns of digits were prepared from series of random numbers, eliminating zero. On a single page only problems of one given length were used. Problems were prepared containing 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, and 25 digits. The digits were typed in primer size type, eight characters per inch, to insure legibility.

Subjects: A total of 193 high-school boys and girls was tested. They were in the age range of 16 to 20 years, were in the fourth year of high school, and were native-born white. They were selected by classes to yield a group of average intelligence using the results of the Otis Self Administering Test given by the school; the mean I.Q. was 104.6, the standard deviation 9.0.

A total of 50 subjects between the ages of 60-69 years was tested. These subjects were convalescing patients in general hospitals, and residents of The Baltimore City Home for the Aged. All subjects were selected after individual interviews in which each subject's vision was tested and his age, education, and nativity were determined.

A total of 33 patients institutionalized because of senile psychoses, i.e., senile psychosis or psychosis with cerebral arteriosclerosis, was tested. The senile patients were selected from three mental institutions with a combined population of about 6000 patients. As in the case of the normal elderly, all senile patients were native-born white with a minimum of four years of education.

In individually administered tests, time to complete the page or a maximum time of two minutes per page was used, whichever came first. In group testing the timing of a page was adjusted by preliminary study so that no individual would complete the page in the time limit. A time limit of one minute was used for the 2, 3, and 4 digit problems, and two minutes for all longer problems; thirty seconds rest was allowed between each page.

Administration: High-school students were given the tests in groups of about 30. All aged subjects were tested individually. In all instances the subjects were told to do the problems as quickly as they could. Rate of uncorrected or total addition is defined as the total

number of addition operations completed divided by the time interval. Rate of correct addition is defined as the number of operations in the problems added correctly divided by the time taken to complete all problems. In group testing the subjects were instructed to place a check mark opposite the digit they were adding when time was called. They were also instructed to place a direction arrow alongside the column indicating the direction in which they added, i.e., top-down or bottom-up. The digits added in the incomplete problems were included in the computation of the rate of addition for the individual. The estimated time spent on the incomplete problem was deducted from the total time in computing the rate of correct addition.

The effects of order of administration of the various lengths of problems were determined by giving the tests in different orders to two groups of 30 high-school senior boys matched for age and I.Q. on the Otis Self Administering Test. One group took the test in a constantly increasing series: 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, and 25 digits. The other group took the tests in the order: 2, 4, 8, 25, 3, 6, 15, 10, 5, 9, and 7. An analysis of variance was made to test the significance of differences in slope and level of performance in the two groups. There were no significant differences in the slope or level of performance when digits added per second was used as the measure of performance. When the logarithm of the operations was used, slight differences appeared which were inconsistent for the analyses of uncorrected addition and correct addition. In the case of uncorrected addition the difference of level was not significant for the two groups and the difference in slope was significant at the 5 per cent level. Less than 0.6 per cent of the sums of squares could be attributed to the differences in slopes between the groups. In the analysis of variance of correct addition, differences in slope were not significant whereas level was significant at the 5 per cent level. These results suggest that order of presentation had little if any effect upon the performance of the subjects.

Scoring: Individual tests were scored both as operations per second and correct operations per second. The results of each individual were graphed in duplicate. Two investigators independently fitted straight lines to the graphs. Results were plotted as log operations per second against log operations in the problem and as log log correct operations per second against operations in the problem. In fitting the straight lines the middle portion of the curve was given emphasis because of the limitation of performance at low difficulty

due to writing speed and the lack of reliability at high difficulty in the elderly.

The correlation between the slopes derived by two investigators was 0.84 for total addition and 0.92 for correct addition in the high-school group. The corresponding correlations were 0.69 and 0.92 for the aged subjects. For each subject the slopes and intercepts derived by the two investigators were averaged. In the analysis of the data, the differences between the groups of subjects were thus secured by the mean of individual rates and not derived from a mean curve.

Not all subjects were used in every comparison. In some instances the subjects were excluded when it was difficult to fit curves to the data because of excessive variability and missing values. In the aged and senile psychotic subjects the total output for long problems per two minutes was in some instances so small as to result in large variability when considering correct addition. During group administration of the tests to high school students an occasional measurement would be missed.

Results

The data show a systematic relation between the rate of addi-

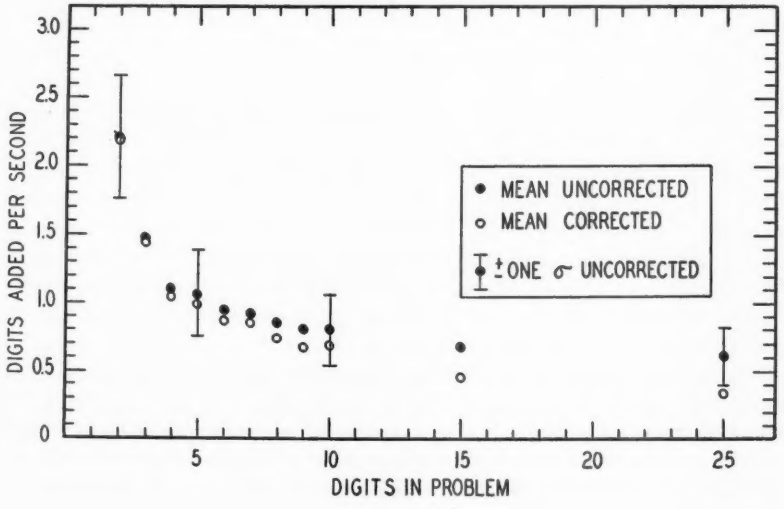


FIGURE 1

Addition rate as a function of problem length. Mean values of 193 senior high-school students age 16 to 20 years are plotted as the mean rate of total uncorrected addition and as mean rate of correct addition.

tion and the length of the problem. However, it is apparent in Figure 1 that this relation is not linear. The relation between rate of addition (not corrected for errors) and problem length was determined to be of the form $Y = \frac{A}{X^n}$; where Y is the rate of addition in operations per unit time, A the initial rate or rate at a difficulty of one operation, n the slope, and X is the number of addition operations per problem.

TABLE 1

Analysis of Variance of Addition Rate Using the Logarithm of Addition Operations

Source	Degrees of Freedom	Sum of Squares	Variance
Total	2100	60.39	
Individuals	190	37.85	0.199
Columns, total	10	13.39	1.339
Linear regression	1	13.04	
Departure from linearity	9	0.35	0.039
Residual	1900	9.15	0.005

Columns: $F = 1.339/0.005 = 267.80$
Curvilinearity: $F = 0.039/0.005 = 7.80$

The goodness of fit (8) of the power function above was determined for the data of the high-school students (Table 1). There is still some slight residual curvilinearity in the data. The equation is accepted as the best representation of the data, however, since the residual curvilinearity is a small portion of the total sum of squares and appears related to the limitation of writing speed in recording the answers to short problems. Small departures of the tests from "true" difficulty at a given level also may contribute to the slight curvilinearity. In short problems a larger portion of the time is spent writing answers so that the total rate of addition appears lowered at the easy end of the difficulty scale. This is shown in Figure 2 wherein the correlations between speed of writing and rate of addition have been correlated for problems of different length. The correlations decline until a stable minimum is reached at problems of five digits or longer. Speed of writing was measured by having the subjects write digits as quickly as they could (1). The correlation between writing speed and rate of addition is lower at every level

for the high-school students than for the sixty-year-old subjects. The suppression of addition rate due to writing speed is apparent in the curves of individual subjects. In Figure 3 the curve of a young adult illustrates the departure from linearity that may occur at low difficulty.

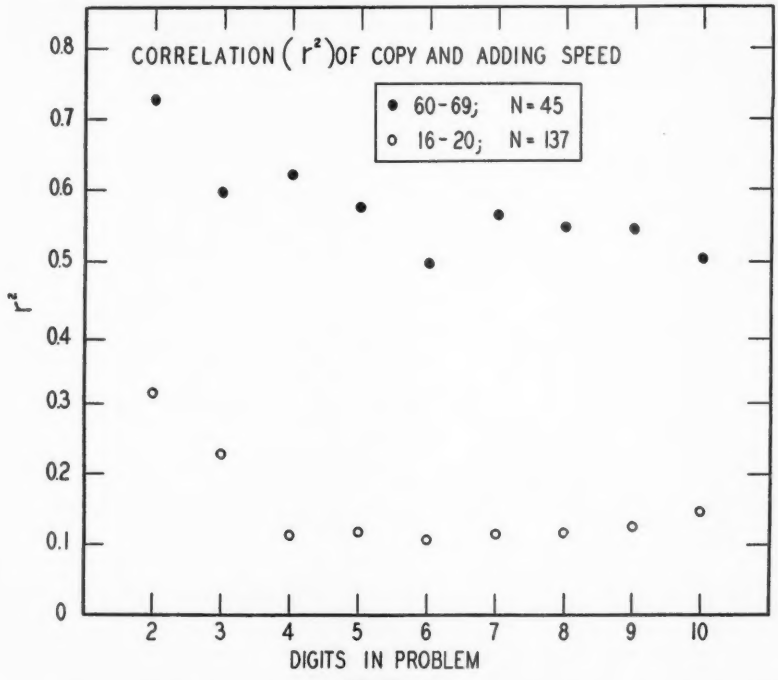


FIGURE 2

Correlation (r^2) between speed of writing digits and the rate of addition for problems of different lengths. Values are based on 137 senior high-school students and on 45 normal elderly subjects.

The dashed line represents an extrapolation of the linear portion of the curve. The Y value when $x = 1$ would represent the theoretical maximum function for this subject, i.e., the rate of addition at a difficulty level of one operation. By graphing an individual subject's performance, the effect of writing speed can thus be excluded from the estimation of the slope and intercept values.

Although the mean curves are only crude representations of the data they suggest that the elderly and the senile subjects show a dif-

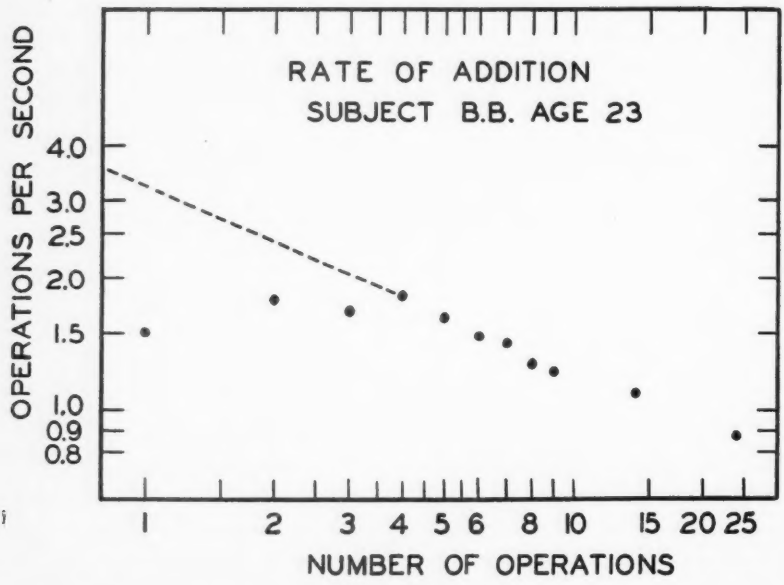


FIGURE 3

Rate of addition as a function of difficulty in a 23 year-old subject. Departure of linearity at low difficulty represents the influence of speed of writing the answers. Extrapolation of the curve yields an estimate of maximum performance unaffected by this disproportionate suppression of addition rate. Subject was given the problems in random order.

ferent level or rate of addition than the high-school students (Figure 4). The exponents of the equations, however, are not significantly different for the three groups. This was demonstrated by comparing the slopes of the individual plots of rate of total addition on log log graph paper (Table 2).

Differences in the performance of the young and aged subjects are more striking in the rate of correct addition. The senile patients appear to decline most rapidly in the rate of correct addition (Figure 5). Thus the senile person not only works more slowly but also less accurately when problem length is increased.

When the rate of addition is corrected for errors in adding, a different equation is required to describe the data. It was found in the individual plots that a linear fit was obtained between the log log rate of correct addition and the number of operations per problem. This indicates that there is an exponential relation between the rate of correct addition and problem length of the form $Y = C^{a/m^b}$; where

TABLE 2
Addition Rate Parameters in Three Classes of Subjects

		High-School Students $N = 184$	Normal elderly $N = 46$	Senile patients $N = 26$
Total	Mean log rate at $X = 1$	1.12	0.98	0.85
	σ	0.14	0.18	0.22
Addition	Mean slope $\frac{\Delta \text{ log rate}}{\Delta \text{ log operations}}$	0.30	0.27	0.35
	σ	0.11	0.10	0.22
Correct	Mean log log rate at $X = 0$	1.04	0.96	0.82
	σ	0.05	0.12	0.26
Addition	Mean slope $\frac{\Delta \text{ log log rate}}{\Delta \text{ operations}}$	0.021	0.023	0.044
	σ	0.013	0.016	0.036
	Mean operations at $Y = 1$	68.0	64.9	40.0
	σ	46.0	44.1	48.0

Extreme skewing is present in some of the measures so the mean and σ values are only rough indications of the differences in the classes of subjects. The X value for $Y = 1$ in rate of total addition was not computed since the groups did not differ in the slope of total addition.

All original values for total addition were multiplied by 10 and all correct addition rates by 100. In securing the antilogarithms of the above values the appropriate divisions should be made.

Y = rate of correct addition. On the log log graphs, log a = ordinate intercept, $-\log m$ = slope, x = number of operations, and C = base of common logarithms. When the slopes of the individual plots were obtained, significant differences were found between the groups (Table 2). A larger difference was noted between the senile and the elderly subjects than between the elderly and the high-school students.

If the X intercept on the log log graph is computed by dividing the Y intercept by the slope for each subject, a value is derived which represents the functional limit for the subjects, i.e., the longest problem that the subject may be expected to solve correctly in unit time (Table 2). This value reveals large differences between the subjects and suggest that the senile patients have suffered a larger per cent

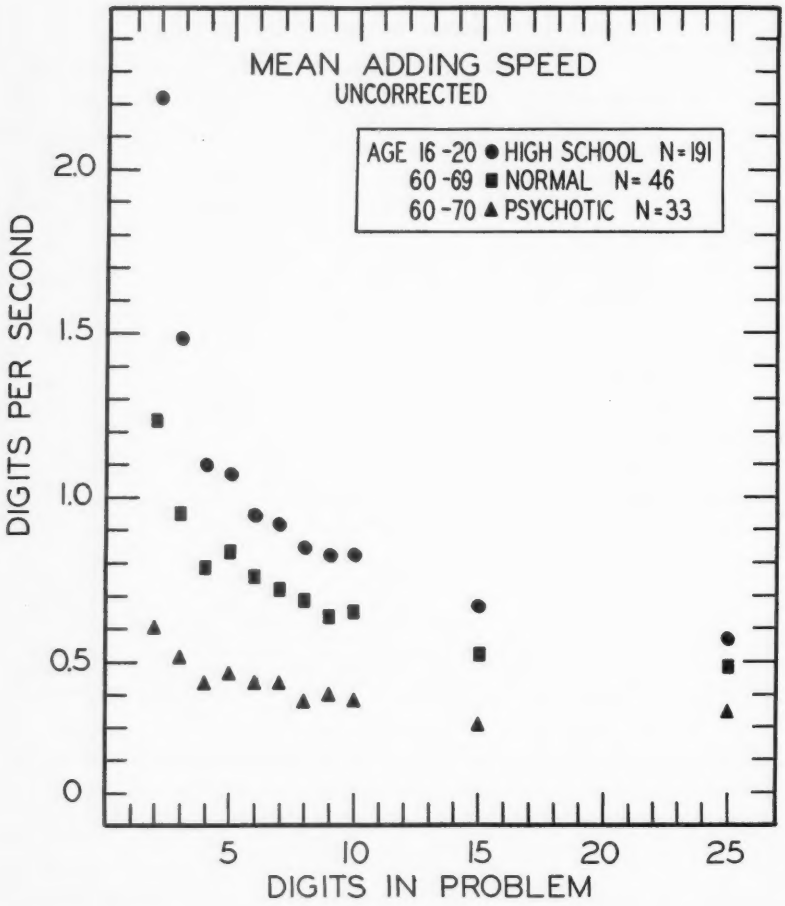


FIGURE 4

Rate of total uncorrected addition as a function of problem length for three classes of subjects: 193 senior high-school students, 46 normal elderly subjects aged 60-69 years, and 33 senile psychotic patients aged 60-70 years.

loss of function than the normal elderly. These values may have biological validity in that they demonstrate that the difference in function between the senile mental patients and the normal individuals of the same age is greater than the difference between the young and elderly normal subjects.

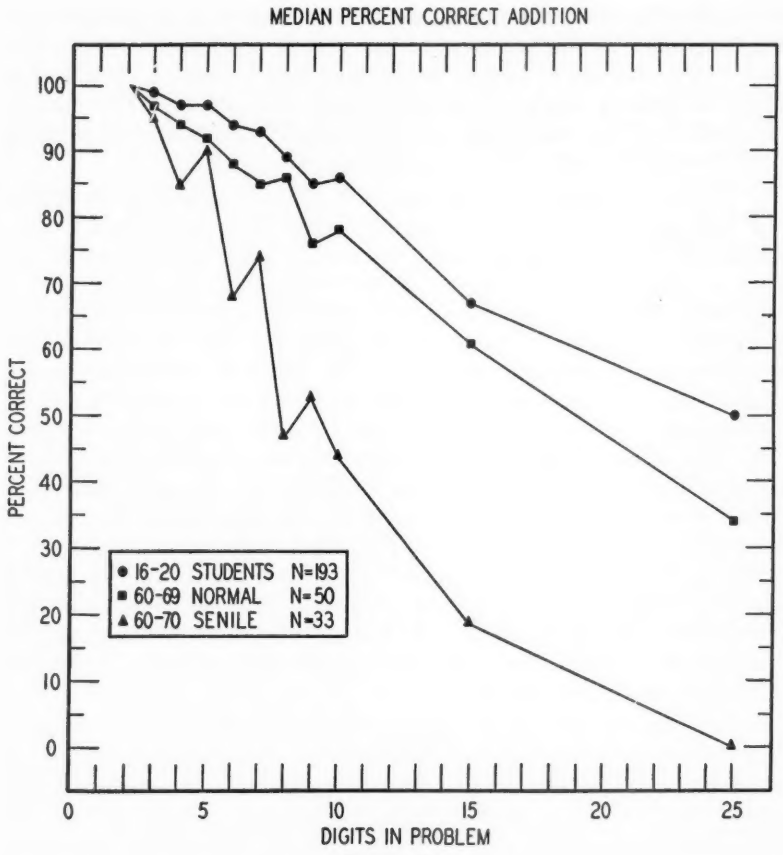


FIGURE 5

Per cent correct addition as a function of problem length. Median values are given for 193 senior high-school students, 50 normal subjects aged 60-69 years and 33 senile mental patients age 60-70 years. Median values were preferred to means because of the frequency of zero correct performances for difficult problems in the aged subjects.

Discussion

The present study has undoubtedly oversimplified many relations and perhaps ignored certain relevant variables. The results clearly suggest, however, that there is value in the approach. The present results provide an empirical background as well as a conceptual framework for future modification. There are several features of these results that have important implications: (1) the

demonstration that rate of addition may be treated as a psychophysical function, (2) the finding that the rate of total and correct addition and the decline of correct addition with increased problem length show differences between young and aged subjects, and (3) the extrapolation of the individual curve permits the extraction of a value representing the subject's theoretical maximum limit.

In the discussion of these results it seems desirable to designate the intercepts and slopes in a way to suggest their conceptual origin. If one has rectified the data and obtained a linear relation as in the case of addition rate, it is simple to derive the parameters of the function. These parameters may be imparted a psychological interpretation. Thus the Y value representing the rate of addition for problems of zero or unit length might be called the *functional maximum,* FM. The X value or problem of maximum length where the rate of addition is zero or unity might be called the *functional limit,* FL. The slope representing the decline in rate with increased difficulty might be called the *functional decline,* FD. The area under the curve represents work completed during the test session and might be called the *functional capacity,* FC. This value was not computed in the present study because in deriving the expression the FM would have to be squared. Thus errors are exaggerated and unless one has considerable confidence that this intercept is accurate it would be best to use the FL as the index of the person's function in comparing him with other individuals.

The slope or FD is presumably free from the effects of individual differences in motivation, criterion of acceptable accuracy, and speed of visual perception and motor response so long as the subject maintains these at a constant level throughout the test series. All other measures, the FM, FL, and the FC are influenced by individual differences in such constants of performance. This would suggest that the FD of the correct addition is more desirable in comparing subjects of widely different ages and background in whom there might be large individual differences in level of motivation. Any transitory states such as warm-up, fatigue, and practice within the test session would influence the slope function, FD. In view of the early work (11), however, it would seem that much longer test sessions than used in this study would be required to elicit sizable effects from such variables.

Whether the results of the present study would apply to tests of double or triple the lengths used in this study is of course not known. It is possible that the functional relation between problem

length and rate of addition would show some deviation from the equations given here for problems of extreme length. The use of problems of a length longer than 25 digits is prohibitive for most of the senile patients and many of the normal elderly. It takes the aged so long to do a problem of 25 digits that it is very difficult to obtain a reliable estimate of their rate of correct addition with paper-and-pencil methods. With this in mind plus the complication of writing speed for short problems, it seems that the optimum length of problems for computing the slope and intercept values for practical applications lie between 5 and 15 digits.

The individual differences in the parameters of rate and level isolated in this study give promise of conceptual value in analyzing the mental performance of different classes of subjects.

Summary and Conclusions

1. The purpose of this study was to examine the relation between the rate of simple addition and the length of the digit series to be added. The hypothesis tested was that the rate of addition would decline in a systematic manner as a function of the logarithm of the number of addition operations per problem. Addition problems were prepared that consisted of random digits arranged in single columns that varied in length from 2 to 10, 15, and 25 digits. The task required the subject to add the digits as quickly as possible. Three groups of subjects varying in age and mental status were used: (A) 193 senior high-school students, (B) 50 subjects aged 60 to 69 years of age, and (C) 33 patients institutionalized for senile psychoses. All subjects were native-born white. The elderly subjects and the senile mental patients were matched for age and education.

2. The rate of addition was found to be a function of the length of problem as measured by the number of operations. The equation expressing this relation for rate of total addition is of the form $Y = \frac{A}{X^n}$. The equation expressing this relation for rate of correct addition is of the form $Y = C^{A/m^s}$. In each case Y is the addition rate in operations per unit time and X the number of operations in the problem.

3. Significant differences were found in the general rate of addition, or value of Y at lowest difficulty, for the three groups of subjects. No difference was noted in the slope of total addition. However, significant differences were found between the groups of subjects in slopes of correct addition. When the Y intercept of correct

addition is divided by the slope to secure the limiting value of X the differences between the groups were accentuated indicating that the senile subjects not only add more slowly but their accuracy drops disproportionately when difficulty is increased. The longest problem they would be expected to do correctly in infinite time was markedly decreased. Because of the results obtained, it is suggested that the parameters derived for each subject be designated in such a way as to suggest their conceptual origin. Thus the value of Y at $X = 0$ or unity is defined as the functional maximum, FM; the slope as the functional decline, FD; the value of X at $Y = 0$ or unity as the functional limit, FL. In addition the area under the curve may be determined and defined as the functional capacity, FC.

4. It has been demonstrated there is a uniform quantitative relation between the rate of addition and the length of the problem to be added. By appropriate graphing of the individual results, estimates of the slope and intercept parameters may be obtained. These parameters appear useful in analyzing the performance of different classes of subjects.

REFERENCES

1. Birren, J. E., and Botwinick, J. The relation of writing speed to age and to the senile psychoses. *J. consult, Psychol.*, 1951, 15, (In Press).
2. Coombs, C. H. The concepts of reliability and homogeneity. *Educ. psychol. Meas.*, 1950, 10, 43-56.
3. Guilford, J. P. The psychophysics of mental test difficulty. *Psychometrika*, 1937, 2, 121-133.
4. Guilford, J. P. The difficulty of a test and its factor composition. *Psychometrika*, 1941, 6, 67-77.
5. Landahl, H. D. Time scores and factor analysis. *Psychometrika*, 1940, 5, 67-74.
6. Oehrn, A. Experimentelle Studien zur Individualpsychologie. *Psychol. Arbeit.*, 1895, 1, 92-151.
7. Richardson, M. W. The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1936, 1, 33-49.
8. Snedecor, G. W. Statistical methods. Ames, Iowa: Iowa State College Press, 1950; pp. 374-399.
9. Tate, M. W. Individual differences in speed of response in mental test materials of varying degrees of difficulty. *Educ. psychol. Meas.*, 1948, 8, 353-374.
10. Thurstone, L. L. Ability, motivation, and speed. *Psychometrika*, 1937, 2, 249-254.
11. Vogt, R. Ueber Ablenkbarkeit und Gewöhnungafähigkeit. *Psychol. Arbeit.*, 1899, 3, 62-201.
12. Whipple, G. M. Manual of mental and physical tests: Part II. Baltimore: Warwick & York, 1915; pp. 460-485.

Manuscript received 9/27/50
Revised Manuscript received 11/13/50

PSYCHOMETRIKA—VOL. 16, NO. 2
JUNE, 1951

A MECHANICAL MODEL ILLUSTRATING THE SCATTER DIAGRAM WITH OBLIQUE TEST VECTORS

HAROLD GULLIKSEN
AND
LEDYARD R TUCKER
PRINCETON UNIVERSITY AND EDUCATIONAL TESTING SERVICE

A mechanical model is described for illustrating changes in the configuration of points when the reference axes are rotated obliquely and the values of the orthogonal projections of the points on the reference axes are maintained constant. This type of transformation is important in factor analysis.

Two different methods of graphic representation are in general use in correlation analysis and factor analysis (1, 2). Since the relations and contrasts between these geometric structures are sometimes rather difficult for students to grasp, a mechanical model was devised to facilitate class explanations. This device is to be described.

In the first of the graphic representations, each *test* is taken as an *orthogonal coordinate axis* and each *person* is represented by a *point* with coordinates equal to his scores on the tests. When there are two tests this results in the conventional scatter diagram. Correlation between the tests results in an elliptical type scatter of points.

In the second method of representation, uncorrelated *factors* or *components* are taken as the *orthogonal coordinate axes.* In the "factor space" both *tests* and *persons* may be represented simultaneously as follows:

(a) each *test,* by a *vector* from the origin with the coordinates of its terminus equal to the *test's factor loadings*;

(b) each *person,* by a point with coordinates equal to *his factor scores.*

This "factor space," containing a *vector* for each *test* and a *point* for each *person,* has several interesting properties. (1) The *test score* of each person may be represented by the perpendicular projection of his point on the test vector. (2) There is a circular, or spherical, configuration of the peoples' points. (3) Correlation between two

tests is indicated by the cosine of the angle between the two test vectors.

The type of transformation between these two systems of geometric representations can be contrasted with the more common "rigid rotation" of axes in which the configuration of points remains constant while the axes are rotated; hence the projections of the points on the axes change. This is a familiar type of transformation which appears in many problems other than those of factor analysis. The change in projections can be represented by a matrix equation. On the other hand, in the type of transformation which we wish to illustrate in this paper there is a change in the angle between axes or vectors representing tests while the configuration of points changes in such a manner that the projections of the points on the test vectors are held constant.* The advantage in factor analysis is that the component elements in the space can be made to take on different significances after this type of geometric transformation.

An example in two-space of the results for the second type of transformation is given in Figure 1. We begin with an initial configuration in which the axes are at 90° separation. The plotted points are assumed to have a positive correlation and hence form an ellipse. If the coordinates of the points are standard scores of individuals on two tests, the correlation between tests is related to the ratio of the principal axes of this ellipse. Now let us consider a transformation such that the coordinate axes are separated by an angle θ. The test scores are still plotted as projections perpendicular to these new axes. An angle θ can be so chosen that the points will form a circle. For this geometric representation the correlation between the two tests is now represented by the cosine of θ. This second configuration is illustrated at the right in Figure 1. The projections of point i are illustrated for each configuration. It is to be noted that the projec-

*Professor Ernst Snapper (Visiting Associate Professor from Southern California, now at the Department of Mathematics, Princeton University) suggested that the change in coordinates of the points on a fixed pair of orthogonal axes can be described as a combination of a stretch transformation and a shear transformation, both parallel to the y-axis. The x coordinates remain constant while the y coordinates undergo the following two transformations:

$$y' = \frac{1}{\cos\theta} y \qquad \text{(stretch transformation)}$$

$$y'' = y' - (\tan\theta)\, x \qquad \text{(shear transformation)}$$

where θ is the angle of rotation of the axis representing test y. The combined transformation is:

$$y'' = \left(\frac{1}{\cos\theta}\right) y - (\tan\theta)\, x.$$

tions on the coordinate axes are identical in the two configurations; however, the point *i* has moved.

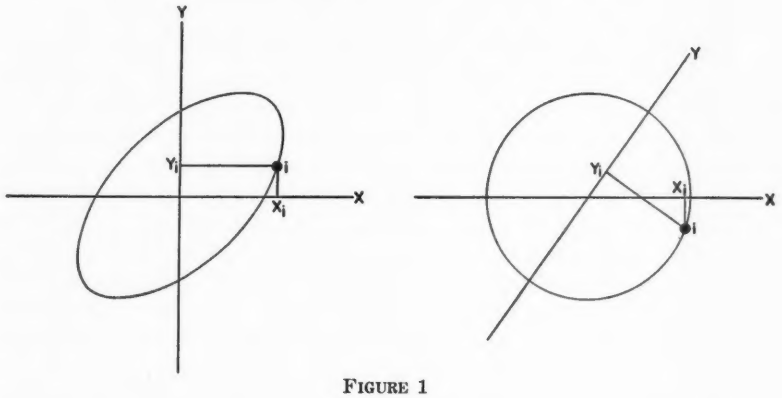


FIGURE 1

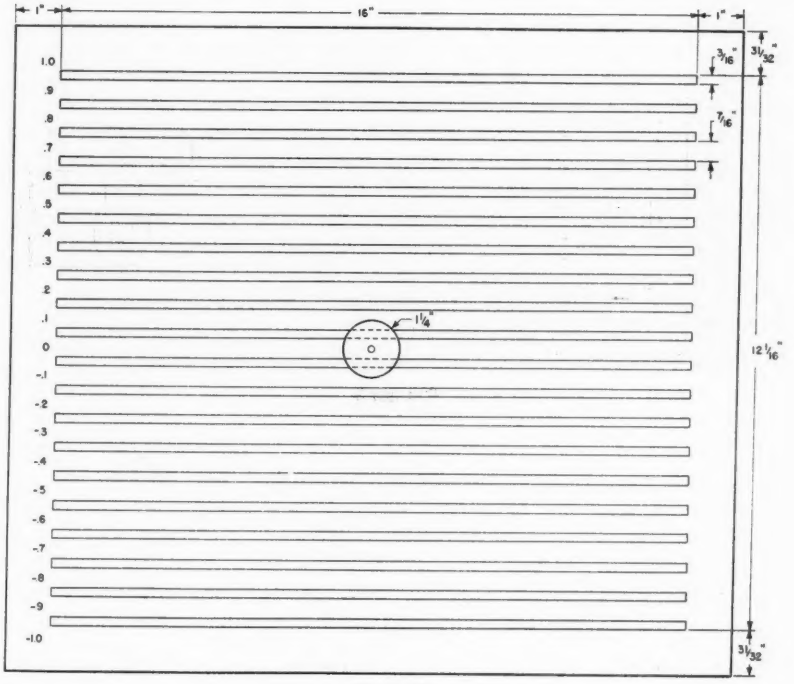


FIGURE 2

This type of transformation is a homogeneous linear non-singular transformation of axes; e.g., a projective transformation of axes. Simultaneously the points in question undergo a coordinate-preserving transformation.

Since students have considerable difficulty in grasping the ideas involved in this second type of transformation, a mechanical device was constructed to illustrate this transformation in a two-space.

The device consists of two boards with slots in them as shown in Figure 2. The bottom board was made of 1/8″ bakelite, 16″ × 20″ (24″ would be better). The top board was smaller. A 14″ × 18″ piece of 3/8″ lucite was used. The boards were pivoted to each other by a bolt at the origin labelled "O". The pivot point was reinforced by a 1 1/4″ metal disk as indicated in the figure. Twenty slots each 3/16″ wide and 16″ long were milled in the boards. These slots were 5/8″ apart (center to center) leaving a 7/16″ strip of material between the slots.

Aluminum pegs, shown in Figure 3, can be inserted at any intersection of two slots, as indicated in Figure 4. These pegs were held in place by snap-in trimounts available from a radio supply house.

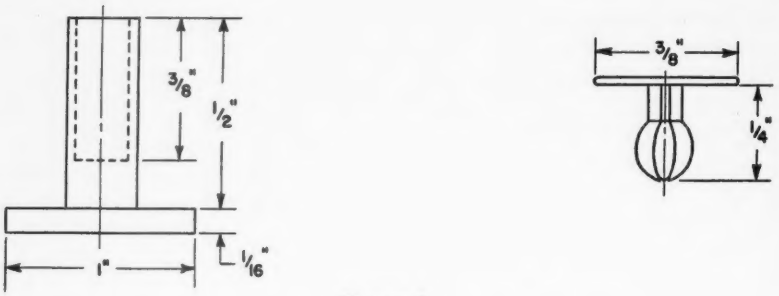


FIGURE 3

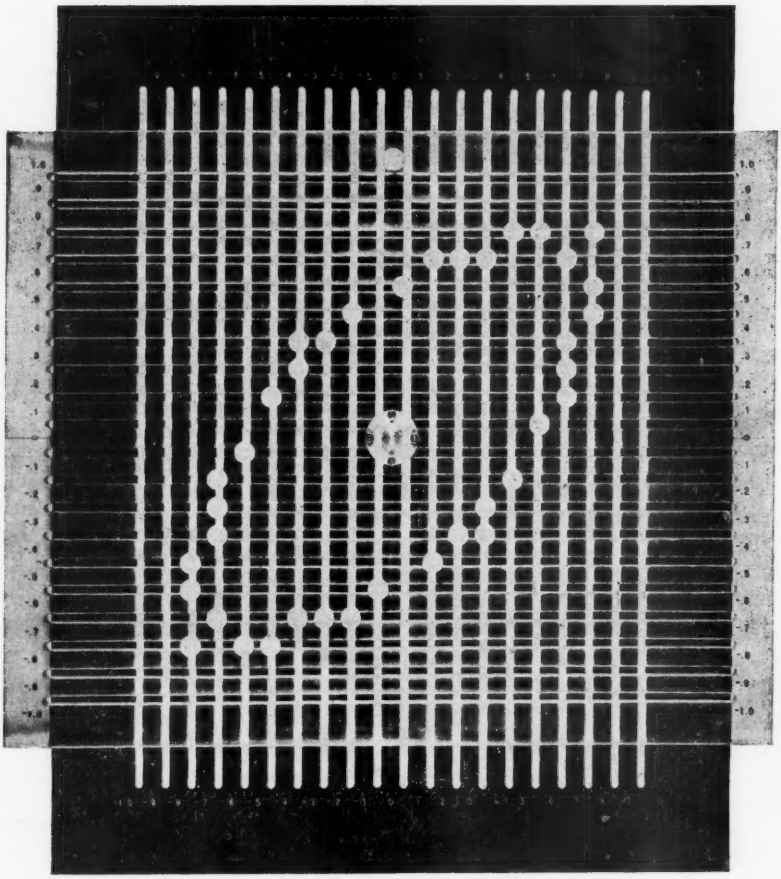


FIGURE 4

The upper lucite board may be rotated from an initial angle of 90 degrees with the lower board to an angle of about 35 degrees. This corresponds to a cosine or "correlation" of about .80. The projection of each point on each axis remains invariant for any rotation but the inter-point distances, or the configuration of points changes as shown in Figure 5. Thus the "correlation" shown by the set of points will be a function of both the original configuration of points and the angle between the axes.

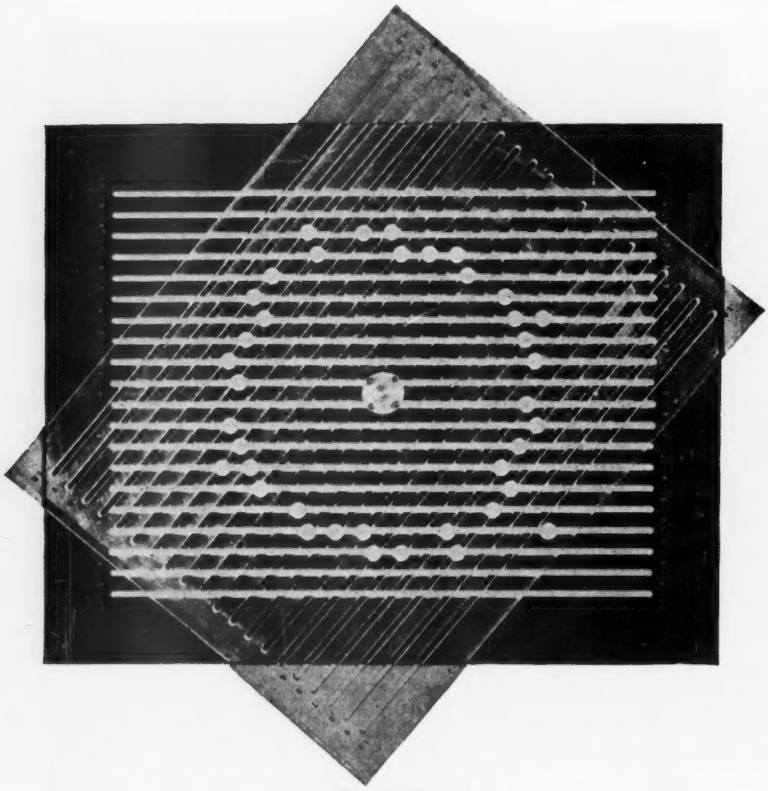


FIGURE 5

With this device any configuration of points can be set up with orthogonal axes and rotated to —35 or + 35 degrees corresponding approximately to correlations of —.8 to +.8. Beyond this rotation, the pegs will tend to jam. This device has furnished a very effective means for demonstrating the effects of transformation from orthogonal to oblique axes for students of factor analysis.

REFERENCES

1. Thomson, Godfrey H. The factorial analysis of human ability. New York: Houghton Mifflin Co., 1946.
2. Thurstone, L. L. Multiple-factor analysis. Chicago: The University of Chicago Press, 1947.

Manuscript received 5/30/50
Revised manuscript received 12/15/50

PSYCHOMETRIKA—VOL. 16, NO. 2
JUNE, 1951

A GRAPHICAL METHOD FOR THE RAPID CALCULATION OF BISERIAL AND POINT BISERIAL CORRELATION IN TEST RESEARCH

HOWARD W. GOHEEN AND MELVIN D. DAVIDOFF
U. S. CIVIL SERVICE COMMISSION

A description is given of a diagram (available separately) for computing biserial or point biserial correlation coefficients. The diagram is maximally useful where large numbers of coefficients are to be calculated in test item analysis. The diagram is entered with the mean criterion score of the group passing the item and the proportion of correct answers to the item.

Current emphasis on test analysis procedure has led to an increased application of biserial and point biserial correlation methods in test-item analysis to which these techniques are particularly adaptable. The nomograph presented on a following page permits particularly rapid calculation of these coefficients with accuracy to the second decimal.* When IBM equipment is available for the determination of item means and proportions, this nomograph represents the shortest method for calculation of these coefficients known to the authors. It is maximally useful where large numbers of coefficients are to be calculated for individual tests. Its application to the calculation of single coefficients from particular data represents little if any advantage over other methods.

Given the mean $(\bar{X})$ and standard deviation (σ) of the total group on the criterion by which the test items are to be evaluated, the only data necessary for determining either of the two coefficients for any item are the proportion of correct answers to that item (p) and the mean criterion score (M) of those who answered the item correctly.

The ordinate markings on the chart are in symbol form so that it can be adapted for use on any particular test. Before using the chart the appropriate values should be entered as indicated along

*The diagram is published in reduced size for illustrative purposes only. Interested readers may obtain a copy of the diagram, size $10\frac{1}{2} \times 16$ inches, without charge by writing to the Test Development Section, United States Civil Service Commission, Washington 25, D. C., requesting a copy of Test Technical Series No.17.

the ordinate of the chart. (Note that these ordinate values will be identical when the chart is turned for computation of biserial r's or for point biserial r's.) If, for example, $\bar{X} = 65$ and $\sigma = 9$, these values would read:

$$\bar{X} + .4\sigma = 68.6$$
$$\bar{X} + .3\sigma = 67.7$$
$$\bar{X} + .2\sigma = 66.8$$
$$\bar{X} + .1\sigma = 65.9$$
$$\bar{X} = \quad 65$$

It will be noted that the ordinate entries simply increase $.1\sigma$ for each value.

Example:

The mean criterion score for the total group ($\bar{X}$) is found to be 65; the standard deviation (σ) of this distribution is 9. Now let

$P_1, P_2, \cdots P_i$ = the proportion answering items 1 through i correctly

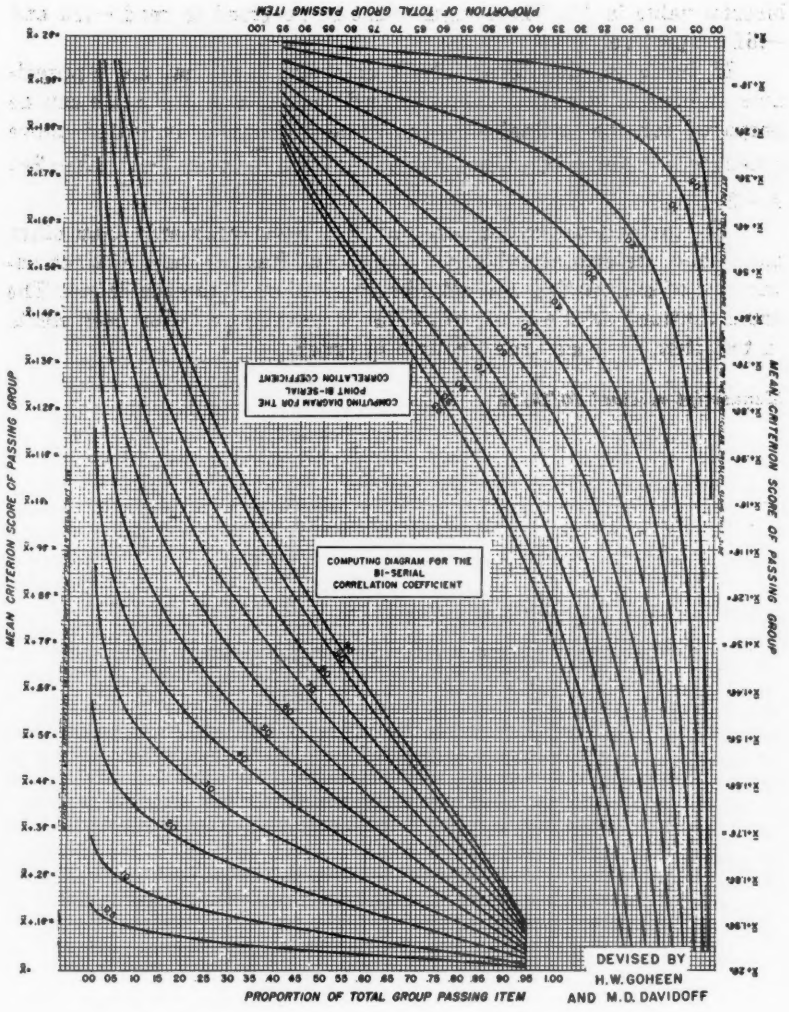
$M_1, M_2, \cdots M_i$ = the mean criterion score of those answering items 1 through i correctly

Suppose that you want the biserial coefficient for item 5 and that 20% of the sample got it right ($p_5 = .20$), and the mean criterion score of those who got it right was 72 ($M_5 = 72$). Find the intersect of 72 on the ordinate and .20 on the abscissa. This yields a biserial r of .55. If the data are such that the point coefficient is the more appropriate, rotate the chart 180°, and read at the intersect as before. For the example given, the point biserial value is .39.

It is apparent that when the mean criterion score value of those passing the item is less than the mean of the total sample the sign of the coefficient will be negative. In such cases it is necessary only to obtain the difference between the mean of the passing group (M) and the mean of the total ($\bar{X}$) and add this difference to the total mean. This can be done by the formula $2\bar{X}-M$. Enter the chart as before, affixing a negative sign to the obtained coefficient.

Example:

Item six yields the following data: $p_6 = .25$; $M_6 = 57$; $2\bar{X} - M = 73$. At the intersect of 73 and .25 the biserial value is .70, the point


COMPUTING DIAGRAM FOR THE BI-SERIAL CORRELATION COEFFICIENT
COMPUTING DIAGRAM FOR THE POINT BI-SERIAL CORRELATION COEFFICIENT
MEAN CRITERION SCORE OF PASSING GROUP
PROPORTION OF TOTAL GROUP PASSING ITEM
DEVISED BY H.W. GOHEEN AND M.D. DAVIDOFF

biserial value is .51. These signs must be reversed to read —.70 and —.51 respectively.

If, however, the item-analysis data are such that any appreciable number of negative values is anticipated, a second scale can be prepared for the ordinates in the same fashion as indicated above except with reversed sign, i.e., ascending from $\overline{X}$ thus: $\overline{X}$—.1σ, $\overline{X}$—.2σ, $\overline{X}$—.3σ, etc.

The method presented here requires fewer constants than other methods of determining biserial values and has the very distinct advantage of eliminating the actual computation of the coefficient. The usual cautions on the interpretations of serial correlation coefficients in test item analysis are of course in order.

Manuscript received 10/19/50

PSYCHOMETRIKA—VOL. 16, NO. 2
JUNE, 1951

BOOK REVIEWS

GEORGE KINGSLEY ZIPF. *Human Behavior and the Principle of Least Effort.* Cambridge: Addison-Wesley Press, Inc., 1949. Pp. xi + 573.

This book touches "economics, sociology, cultural anthropology, psychology —both general and Freudian—linguistics, and semantics" by the author's own admission. The author's thesis is that, in order to survive, an organism must select a course of action which will minimize its present work plus its probable work of the future. This is referred to as the *principle of least effort.* Zipf applies this principle to virtually all behavior of all organisms. In the linguistic area, he considers such items as frequency and rank order of words, word meanings and rank, frequency and number of words, and repetition interval and number of words, in connection with samples of the speech and written products of children, adults, and psychotics, in many different languages. The sizes of cities, number of retail stores, factories, service and business establishments, news items in daily papers, obituaries in the *Times,* marriage licenses issued in Philadelphia as a function of the blocks separating the couples, and the distribution communities in different nations all lend support for the *least effort* principle according to Zipf.

Many interesting empirical curves are presented, and the author shows great originality on every page; but the reviewer, after long and painful effort, could not follow the author's logic or mathematics in relating the curves to the hypothesis.

At times it is extremely difficult to take Zipf seriously. For example, the author can make procreation fit into the least effort doctrine only by assuming an ego or "identity-point" which survives the organism. He admits that he doesn't know whether "it" is or is not eternal, and if "it" survives death he doesn't know what "it" is. But it leads him to the conclusion that there is always the same fixed number of organisms alive on the planet at any one time. The reviewer is aware of a number of lines of reasoning which have led men to posit an eternal soul, but this is the only soul, to his knowledge, which is required by virtue of the general fact of procreation. Another indication of Zipf's line of attack is the following: '. . . For as soon as the father falls in love with a child and "wants" or "desires" it, he polarizes himself in the opposite direction sexually, and thereby theoretically sets up reactions to produce a child of the opposite sex, thus confirming the hoary superstition that the sex of one's offspring is the opposite from that desired by the parents,' p. 262. Zipf accomplishes similar feats on every other page.

Zipf has been publishing material of this sort since 1929. His present work represents the broadest and most detailed application of his approach. This reviewer feels it cannot be taken seriously as a scientific effort.

University of Wisconsin *David A. Grant*

ALPHONSE CHAPANIS, WENDELL R. GARNER, CLIFFORD T. MORGAN. *Applied Experimental Psychology*. New York: John Wiley & Sons, Inc., 1949. Pp. xi + 434.

Applied Experimental Psychology is an outgrowth of a series of lectures presented by the authors to engineering students at the Naval Post Graduate School, Annapolis, in the Spring of 1947. The lectures were subsequently printed as a classified Navy publication and now in this form.

The subject matter of the book has been referred to variously as human engineering, engineering psychology, biomechanics, applied psychophysics, psychotechnology, and systems research. Whatever its proper name, the discipline is that which was so stimulated by the recent war. It proceeds from the basic tenet that machines (and systems of men and machines) must be designed in terms of the human abilities and limitations of the operators.

This book will be of great interest to engineers, technicians, and industrial designers; it will acquaint these people with the activities of psychologists; it gives concrete examples of the use of psychological research results in design work. *Applied Experimental Psychology* will do a good public relations job for professional psychology because of the authors' serious effort, by and large successful, to describe research methods and results in terms that can be understood by non-psychologists. Their example of the grouping of production data into analysis of variance tables shows clearly, without going into theory or computing formulas, the idea behind the technique. There is a commendable emphasis on methodology which should better explain the tools that the psychologist brings to bear on problems he tackles.

Which is not to say that *Applied Experimental Psychology* is not an important contribution to psychology itself. It is a first statement of the structure and content of this aspiring discipline. As such, this book must be considered as a text for a course which now must be an integral part of the curriculum of industrial psychology, and, it is to be hoped, for engineering as well.

The text is divided into four sections: methodology, sensation and perception, motor behavior, and the working environment. The authors do not imply by word or tone that this structure has been completely or finally filled in; they do, in fact, point out the gaps and areas of controversy. But in discussing the controversies, they are willing, as applied scientists, to render judgments for the benefit of those who are more concerned with positive suggestions than the recitation of theoretical conflicts.

Not the least of the values of *Applied Experimental Psychology* is in setting the tone for effective employment of the applied scientist. It has been and still is difficult for the applied scientist, especially the psychologist, to find his most meaningful mode of activity. There are all too few examples of practicing psychologists who are able, in the competitive industrial world, to utilize in a practical way the best of academic techniques and attitudes in finding the answers to industrial problems. This text illustrates the use of fundamental research techniques in a powerful way for the solution of pressing questions of the day. In prosecution of this argument, it must be said that *Applied Experimental Psychology* exposes a number of superficial treatments of data. Research on systems is an area where only the most cursory techniques have been developed. There is a challenge here for the psychometrician. Considerably ingenuity of the highest order is required of the applied scientist in discovering molar concepts

and measuring them under conditions which are not as convenient as the laboratory. Incomplete data and data which are biased by known variables can and must be dealt with by an applied psychologist in a competent way. The alternative, so frequently invoked, is the "quick and dirty" job, whose ultimate value for both the consumer and the applied scientist is questionable.

Applied Experimental Psychology is, then, an estimable effort. It has immediate value for the industrial designer, the industrial psychologist, students of both industrial psychology and engineering, and for the profession of psychology at large.

New York University *Robert L. Chapman*

BOOKS RECEIVED

CALVIN P. STONE (Ed.) *Annual Review of Psychology*, Vol. II. Stanford: Annual Reviews, Inc., 1951. Pp. ix + 389.

Tables d' Interets et d' Annuites. Brussels: Credit Communal De Beligique, 1950.